Universidade de Aveiro Departamento de Química
2015

Diana Maria de Figueiredo Pinto

Marcadores moleculares para a Nefropatia Diabética

Molecular markers for Diabetic Nephropathy

Universidade de Aveiro Departamento de Química
2015

Diana Maria de Figueiredo Pinto

Marcadores moleculares para a Nefropatia Diabética

Molecular markers for Diabetic Nephropathy

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Bioquímica, ramo de Bioquímica Clínica, realizada sob a orientação científica da Doutora Maria Conceição Venâncio Egas, investigadora do Centro de Neurociências e Biologia Celular da Universidade de Coimbra, e da Doutora Rita Maria Pinho Ferreira, professora auxiliar do Departamento de Química da Universidade de Aveiro.

Este trabalho foi efetuado no âmbito do programa COMPETE, através do projeto DoIT – Desenvolvimento e Operacionalização da Investigação de Translação, ref: FCOMP-01-0202-FEDER-013853.

o júri

presidente — Prof. Francisco Manuel Lemos Amado
professor associado do Departamento de Química da Universidade de Aveiro

Doutora Maria do Rosário Pires Maia Neves Almeida
investigadora do Centro de Neurociências e Biologia Celular da Universidade de Coimbra

Doutora Maria Conceição Venâncio Egas
investigadora do Centro de Neurociências e Biologia Celular da Universidade de Coimbra

Agradecimentos

Em primeiro lugar quero expressar o meu agradecimento à Doutora Conceição Egas, orientadora desta dissertação, pelo seu apoio, palavras de incentivo e disponibilidade demonstrada em todas as fases que levaram à concretização do presente trabalho. Obrigada pelo saber transmitido, que tanto contribuiu para elevar os meus conhecimentos científicos, assim como pela oportunidade de integrar o seu grupo de investigação. O seu apoio e sugestões foram determinantes para a realização deste estudo.

À Doutora Rita Ferreira, o meu sincero agradecimento pela co-orientação desta dissertação, pelas suas sugestões e comentários, bem como pela disponibilidade que sempre demonstrou ao longo deste trabalho.

A todas as pessoas da Genoinseq, o meu muito obrigada pela forma como fui recebida, pelo permanente apoio e pelo espírito de entreajuda que sempre esteve presente. Agradeço, de forma especial, à Susana Carmona, ao Felipe Santos, à Joana Sousa e ao Abel Sousa por toda a boa disposição, companheirismo e estímulo nas alturas de desânimo.

Por último, mas não menos importante, dirijo um agradecimento a toda a minha família, em especial aos meus pais e namorado, pelo apoio incondicional, incentivo e paciência constante. Obrigada por toda a ajuda na superação dos obstáculos que ao longo deste percurso foram surgindo e por sempre acreditarem em mim.

Palavras-Chave Diabetes tipo 2, nefropatia diabética, sequenciação de exomas, sequenciação massiva paralela, marcadores moleculares

Resumo

A diabetes tipo 2 é um dos distúrbios metabólicos mais comuns no mundo. Globalmente, está previsto um aumento da sua prevalência, assim como um aumento do risco de desenvolver complicações associadas. Uma dessas complicações é a nefropatia diabética, definida pelo aumento progressivo de proteinúria e um declínio gradual da função renal. Aproximadamente 25% a 30% dos indivíduos com diabetes tipo 2 desenvolvem esta complicação. No entanto, os mecanismos genéticos associados permanecem por esclarecer. Posto isto, o objetivo deste estudo é contribuir para a identificação dos mecanismos envolvidos no desenvolvimento e progressão desta complicação, através da identificação de variantes genéticas relevantes, em indivíduos com diabetes tipo 2 na população portuguesa.

Para isso, os exomas de 36 portugueses com diabetes tipo 2 foram sequenciados na plataforma Ion Proton^{TM}. Desses individuos, 19 não apresentavam nefropatia diabética, tendo sido incluídos no grupo de controlo, e os restantes 17 individuos, com a complicação diagnosticada, formaram o grupo dos casos. Uma análise estatística foi depois realizada para identificar, com base nas diferenças genéticas entre os dois grupos, variantes comuns, assim como genes que acumulam variantes raras candidatas, que podem explicar o risco acrescido ou diminuído para desenvolver a complicação.

Na pesquisa das variantes comuns, as variantes rs1051303 e o rs1131620 no gene *LTBP4*, a variante rs660339 no *UCP2*, a variante rs2589156 no gene RPTOR, a variante rs2304483 no *SLC12A3* e, por fim, a variante rs10169718 presente no gene *ARPC2*, foram, de todas aquelas consideradas estatisticamente significativas (p-value $\leq$ 0,05), as mais relevantes para a patogénese da nefropatia diabética. O rs1051303 e o rs1131620, assim como o rs660339 e o rs2589156, têm um efeito protetor, enquanto o rs2304483 e o rs10169718 foram considerados de risco, estando associados a indivíduos que sofrem da complicação referida. Pela abordagem utilizada para identificar as variantes raras, o gene *STAB1*, que acumula 9 variantes, e o gene *CUX1*, que acumula 2, foram, de todos os genes com significado estatístico (p-value $\leq$ 0,05), aqueles que se evidenciaram como sendo biologicamente relevantes. Ambos os genes foram considerados protetores, já que as suas variantes raras acumuladas estavam presentes maioritariamente nos indivíduos que não apresentam esta complicação renal.

Este estudo providencia uma análise inicial das evidências genéticas associadas ao desenvolvimento e progressão da nefropatia diabética, podendo os seus reultados contribuir para uma melhor compreensão dos mecanismos genéticos que estão por detrás do seu surgimento.

Keywords Type 2 diabetes, diabetic nephropathy, whole-exome sequencing, next-generation sequencing, molecular markers

Abstract

Type 2 diabetes is one of the most common metabolic disorders in the world. Globally, the prevalence of this disorder is predicted to increase, along with the risk of developing diabetic related complications. One of those complications is diabetic nephropathy, defined by a progressive increase in proteinuria and a gradual decline in renal function. Approximately 25% to 30% of type 2 diabetic individuals develop this complication. However, its underlying genetic mechanisms remain unclear. Thus, the aim of this study is to contribute to the discovery of the genetic mechanisms involved in the development and progression of diabetic nephropathy, through the identification of relevant genetic variants in Portuguese type 2 diabetic individuals.

The exomes of 36 Portuguese type 2 diabetic individuals were sequenced on the Ion Proton™ Sequencer. From those individuals, 19 did not present diabetic nephropathy, being included in the control group, while the 17 individuals that presented the diabetic complication formed the case group. A statistical analysis was then performed to identify candidate common genetic variants, as well as genes accumulating rare variants that could be associated with diabetic nephropathy.

From the search for common variants in the study population, the statistically significant (p-value ≤ 0.05) variants rs1051303 and rs1131620 in the *LTBP4* gene, rs660339 in *UCP2*, rs2589156 in RPTOR, rs2304483 in the *SLC12A3* gene and rs10169718 present in *ARPC2*, were considered as the most biologically relevant to the pathogenesis of diabetic nephropathy. The variants rs1051303 and rs1131620, as well as the variants rs660339 and rs2589156 were associated with protective effects in the development of the complication, while rs2304483 and rs10169718 were considered risk variants, being present in individuals with diagnosed diabetic nephropathy. In the rare variants approach, the genes with statistical significance (p-value ≤ 0.05) found, the *STAB1* gene, accumulating 9 rare variants, and the *CUX1* gene, accumulating 2 rare variants, were identified as the most relevant. Both genes were considered protective, with the accumulated rare variants mainly present in the group without the renal complication.

The present study provides an initial analysis of the genetic evidence associated with the development and progression of diabetic nephropathy, and the results obtained may contribute to a deeper understanding of the genetic mechanisms associated with this diabetic complication.

General Index

Figures Index

Tables Index

Abbreviations and Symbols Index

8-Cys	8-Cysteine
1000G	1000 Genomes Project
A	Adenine
ACE	Angiotensin-Converting Enzyme
Acyl-CoA	Acyl-Coenzyme A
ADP	Adenosine Diphosphate
AGER	Advanced Glyction End-Product Receptor
AGE-R1	Advanced Glyction End-Product Receptor 1
AGE-R2	Advanced Glyction End-Product Receptor 2
AGE-R3	Advanced Glyction End-Product Receptor 3
AGEs	Advanced Glycation End-Products
AKT	Alpha serine/threonine Protein Kinase
Alt.	Altered
AMPK	Adenosine 3′,5′-Monophosphate-activated Protein Kinase
Ang II	Angiotensin II
ANP	Atrial Natriuretic Peptide
AP-1	Activator Protein-1
AR	Aldose Reductase
ARBs	Angiotensin Receptor Blockers
ARP2	Actin-Related Protein 2
ARP3	Actin-Related Protein 3
ARP2/3	Actin-Related Protein-2/3 complex
ARPC1	Actin-Related Protein Complex 1
ARPC2	Actin-Related Protein Complex 2
ARPC3	Actin-Related Protein Complex 3
ARPC4	Actin-Related Protein Complex 4
ARPC5	Actin-Related Protein Complex 5
ASO-PCR	Allele-Specific Oligonucleotide Polymerase Chain Reaction
ATP	Adenosine Triphosphate
BAM	Binary Alignment/Map
BMPs	Bone Morphogenic Proteins

bp	Base-pairs
BWA	Burrows-Wheeler Aligner
C	Cytosine
ºC	Degrees Celsius
C^-	Negative Control
Ca^{2+}	Calcium Ion
$[Ca^{2+}]i$	Intracellular concentration of Calcium
CADD	Combined Annotation Dependent Depletion
CBF	CCAAT Binding Factor
CC	Coiled-Coil
CCDS	Collaborative Consensus Coding Sequence
CD36	Cluster of Differentiation 36
CD2AP	CD2-Associated Protein
CDK5	Cyclin-Dependent Kinase 5
ChIP-Seq	Chromatin Immunoprecipitation Sequencing
Chr	Chromosome
CI	Confidence Interval
CKD	Chronic Kidney Disease
Cl^-	Chloride
ClinVar	Clinical Variation Database
CML	N(6)-Carboxymethyllysine
CNV	Copy Number Variation
Co-Smads	Common-partner Smads
COX	Cyclooxygenase
CR1	Cut Repeat 1
CR2	Cut Repeat 2
CR3	Cut Repeat 3
CTGF	Connective Tissue Growth Factor
Cys	Cysteine
Cyt c	Cytochrome c
DAG	Diacylglycerol
dATP	Deoxyadenosine Triphosphate

dbSNP	Single Nucleotide Polymorphism Database
dCTP	Deoxycytidine Triphosphate
ddNTPs	Dideoxynucleotides
DEL	Deletion
DEL HM	Homozygous Deletion
DEL HT	Heterozygous Deletion
Deptor	DEP-domain containing Mammalian Target of Rapamycin-interacting Protein
dGTP	Deoxyguanosine Triphosphate
DM	Diabetes Mellitus
DMD	Duchenne Muscular Dystrophy
DNA	Deoxyribonucleic Acid
dNTPs	Deoxynucleotides
dTTP	Deoxythymidine Triphosphate
ECM	Extracellular Matrix
EGF	Epidermal Growth Factor
ENCODE	Encyclopedia of Deoxyribonucleic Acid Elements
EPACTS	Efficient and Parallelizable Association Container Toolbox
ePCR	Emulsion Polymerase Chain Reaction
EPO	Erythropoietin
ER	Endoplasmic Reticulum
ESP	Exome Sequencing Project
ESRD	End-Stage Renal Disease
ET-1	Endothelin-1
EUR	Europe
$FADH_2$	Flavin Adenine Dinucleotide
FAS1	Fasciclin I
FAT1	FAT Atypical Cadherin 1
FAT2	FAT Atypical Cadherin 2
FFA	Free Fatty Acids
Frag.	Fragment
FSGS	Focal Segmental Glomerulosclerosis
g	Grams

G	Guanine
GBM	Glomerular Basememt Membrane
GC	Guanine-Cytosine
GCS	Gene Candidate Studies
GEMINI	Genome Mining
GERP	Genomic Evolutionary Rate Profiling
GFR	Glomerular Filtration Rate
GLUT2	Glucose Transporter type 2
GLUT4	Glucose Transporter type 4
GOLD	Glyoxal-Lysine Dimer
GPI	Glycosylphosphatidylinositol
Grb2	Growth Factor Receptor-bound Protein 2
GSH	Glutathione
GTP	Guanosine Triphosphate
GWAS	Genome-Wide Association Studies
H	Hydrogen Atom
H^+	Hydrogen Ion
H_2O	Water
HbA1c	Glycated Hemoglobin
HD	Homeodomain
HDIV	HumDiv
HGMD	Human Gene Mutation Database
HM	Homozygous
HM Alt.	Altered Homozygous
$HO_2^{\cdot}$	Hydroperoxyl Radical
HPRD	Human Protein Reference Database
HSF	Human Splicing Finder
HSPs	Heat-Shock Proteins
HT	Heterozygous
HVAR	HumVar
HWE	Hardy-Weinberg Equilibrium
IBM	International Business Machines Corporation

IFG	Impaired Fasting Glucose
IGT	Impaired Glucose Tolerance
IL-1	Interleukin-1
IL-6	Inteleukin-6
IL-10	Interleukin-10
In	Autoinhibitory Domain
Indels	Insertions/Deletions
INS	Insertion
INS HM	Homozygous Insertion
INS HT	Heterozygous Insertion
INSR	Insulin Receptor
IRS-1	Insulin Receptor Substrate 1
IRS-2	Insulin Receptor Substrate 2
ISFET	Ion Sensitive Field Effect Transistor
I-Smads	Inhibitory Smads
ISPs	Ion Sphere^{TM} Particles
K_{ATP}	Adenosine Triphosphate-dependent Potassium Channel
kb	Kilobase
kDa	Kilodalton
KEGG	Kyoto Encyclopedia of Genes and Genomes
kg	Kilograms
LA	Linkage Analysis
LAP	Latency-Associated Peptide
LTBPs	Latent Transforming Growth Factor β Binding Protein
M	Molar
MA	Meta-Analysis
MAF	Minor Allele Frequency
MAPK	Mitogen-Activated Protein Kinase
MCP-1	Monocyte Chemotactic Protein-1
MeDIP-Seq	Methylated Deoxyribonucleic Acid Immunoprecipitation
mg	Milligrams
MG	Methylglyoxal

$MgCl_2$	Magnesium Chloride
miRNA	Micro Ribonucleic Acid
mLST8	Mammalian Lethal with Sec13 Protein 8
mM	Millimolar
mmol/L	Millimoles per Liter
MMPs	Matrix Metalloproteinases
MNP HM	Homozygous Multiple Nucleotide Polymorphism
MNP HT	Heterozygous Multiple Nucleotide Polymorphism
MNPs	Multiple Nucleotide Polymorphisms
Mn-SOD	Manganese-Dependent Superoxide Dismutase
mRNA	Messenger Ribonucleic Acid
MT-MMPs	Membrane-Type Matrix Metalloproteinases
mTOR	Mammalian Target of Rapamycin
mTORC1	Mammalian Target of Rapamycin Complex 1
mTORC2	Mammalian Target of Rapamycin Complex 2
Na^+	Sodium
Na^+/Cl^-	Sodium/Chloride
NADH	Nicotinamide Adenine Dinucleotide
NADPH	Nicotinamide Adenine Dinucleotide Phosphate
Nck	Non-Catalytic Region of Tyrosine Kinase Adaptor Protein
Neph1	Nephrin-like Protein 1
Neph2	Nephrin-like Protein 2
NF	Nuclear Factor
NF-kB	Nuclear Factor kappa B
ng	Nanograms
ng/µL	Nanograms per Microliter
NGS	Next-Generation Sequencing
NH_4	Ammonium
nm	Nanometers
NNSplice	Splice Site Prediction by Neural Network
NO	Nitric Oxide
NPF	Nucleation-Promoting Factor

N-WASP	Neural Wiskott-Aldrich Syndrome Protein
$O_2^{\cdot -}$	Superoxide
OH	Hydroxide
OR	Odds Ratio
PacBio	Pacific Biosciences RS
P-Cadherin	Placental Cadherin
PCR	Polymerase Chain Reaction
P_{GC}	Glomerular Capillary Hydraulic Pressure
PGM	Personal Genome Machine
pHFET	pH-Sensitive Field Effect Transistor
Pi	Inorganic Phosphate
PI3K	Phosphatidylinositol 3-Kinase
PKC	Protein Kinase C
pM	Picomolar
PM	Plasma Membrane
PolyPhen-2	Polymorphism Phenotyping v2
PP	Pancreatic Polypeptide
PRAS40	Proline-Rich AKT Substrate 40 kilodalton
PRCGXPD	Proline-Arginine-Cysteine-Glycine-X-Proline-Aspartate motif
PT	Portugal
Q	Coenzyme Q
R1	Repression Domain 1
R2	Repression Domain 2
RAAS	Renin-Angiotensin-Aldosterone System
RAF	Risk Allele Frequency
RBP4	Retinol-Binding Protein 4
Ref.	Reference
Rheb	Ras Homolog Enriched in Brain
R-I	Type I Receptor
R-II	Type II Receptor

Rictor	Regulatory-Associated Protein of Mammalian Target of Rapamycin, Complex 1 Independent Companion of Mammalian Target of Rapamycin, Complex 2
RNA	Ribonucleic Acid
RNA-Seq	Ribonucleic Acid Sequencing
ROS	Reactive Oxygen Species
RPTOR	Regulatory-Associated Protein of Mammalian Target of Rapamycin, Complex 1
RRP	Readily Releaable Pool
R-Smads	Receptor-Regulated Smads
RXR/KR	Arginine-X-Arginine/Lysine-Arginine motif
SAM	Sequence Alignment/Map
SARA	Smad-Anchor for Receptor Activation
SIFT	Sorting Intolerant From Tolerant
SLR Proteoglycans	Small Leucine-Rich Proteoglycans
SMAD-P	Smad Phosphorylation
SNAP23	Synaptosomal-Associated Protein 23 kilodalton
SNAP25	Synaptosomal-Associated Protein 25 kilodalton
SNAP 23/25	Synaptosomal-Associated Protein 23/25 kilodalton
SNARE	Soluble N-Ethylmaleimide-sensitive Factor Activating Protein Receptor
SNP HM	Homozygous Single Nucleotide Polymorphism
SNP HT	Heterozygous Single Nucleotide Polymorphism
SNPs	Single Nucleotide Polymorphisms
SNV	Single Nucleotide Variant
SOD	Superoxide Dismutase
Solcar	Solute Carrier
SOLiD	Sequencing by Oligo Ligation Detection
SPSS	Statistical Package for the Social Sciences
SR-AI	Scavenger Receptor class A type I
SR-AII	Scavenger Receptor class A type II
SR-BI	Scavenger Receptor class B type I
SR-BII	Scavenger Receptor class B type II
STAB1	Stabilin-1

STAB2	Stabilin-2
T	Thymine
T1D	Type 1 Diabetes
T2D	Type 2 Diabetes
Ta	Annealing Temperature
TAE	Tris-Acetate-Ethylenediamine Tetraacetic Acid
TD-PCR	Touchdown Polymerase Chain Reaction
TGF	Transforming Growth Factor
TGF-β	Transforming Growth Factor β
TIMPs	Tissue Inhibitor of Metalloproteinases
TMAP	Torrent MAPper Aligner
TNF	Tumor Necrosis Factor
TNFα	Tumor Necrosis Factor α
TSC1	Tuberous Sclerosis Complex 1
TSC2	Tuberous Sclerosis Complex 2
TSS	Transcriptional Start Site
TVC	Torrent Variant Caller
U/µL	Units per Microliter
UAE	Urinary Albumin Excretion
UCPs	Uncoupling Proteins
UniProtKB	Universal Protein KnowledgeBase
UTR	Untranslated Region
V	Volts
VAMP2	Vesicle-Associated Membrane Protein 2
VCF	Variant Call Format
VDCC	Voltage-Dependent Calcium Channel
VEGF	Vascular Endothelial Growth Factor
VEGFR	Vascular Endothelial Growth Factor Receptor
VEP	Variant Effect Predictor
WES	Whole-Exome Sequencing
WGS	Whole-Genome Sequencing
WHO	World Health Organization

WT	Wild-Type
w/v	Weight/Volume
Zn	Zinc Ion
ZO-1	Zonula Occludens-1
µg/mL	Micrograms per Milliliter
µL	Microliter
µM	Micromolar
χ^2	Chi-Square

CHAPTER 1| Introduction

1. Diabetes Mellitus

Diabetes Mellitus (DM), often simply called diabetes, is a metabolic syndrome characterized by abnormally high blood glucose levels (hyperglycemia). The most common forms of diabetes are type 1 diabetes (T1D) and type 2 diabetes (T2D). T2D is the most prevalent form of diabetes affecting 90% of the diabetic population (1,2). Both types of diabetes can lead to hyperglycemia, excessive urine production, compensatory thirst and increased fluid intake, among others (1).

Diabetes is diagnosed according to the World Health Organization (WHO) 1999 recommendations (3) (Table 1). Hyperglycemia is measured by the blood glucose level after a minimum 8 hours fasting and 2 hours upon glucose load (75 grams (g)). The glucose concentration measured will determine if a person has pre-diabetes, defined as impaired glucose tolerance (IGT) and/or impaired fasting glucose (IFG), or DM (4).

Table 1. Diabetes diagnostic criteria and other categories of hyperglycemia (adapted from (4)).

	Glucose concentration in venous plasma (mmol/L)
Diabetes Mellitus	Fasting ≥ 7.0 or 2 hour post-glucose load ≥ 11.1
Impaired glucose tolerance	Fasting (if measured) < 7.0 and 2 hour post-glucose load ≥ 7.8 and < 11.1
Impaired fasting glucose	Fasting ≥ 6.1 and < 7.0 and 2 hour post-glucose load (if measured) < 7.8

mmol/L: millimoles per liter.
Glucose load – 75 g glucose orally

1.1. Type 2 Diabetes

Type 2 diabetes is one of the most common metabolic disorders in the world (5). Globally the prevalence of diabetes has been increasing at an alarming rate in the last 15 years, being estimated that 381,800,000 people worldwide (8.3%) currently have this disorder. This number is expected to rise to 591,900,000 people, implying that there will be

a 55% increase in diabetes by 2035 (5). The estimates for the number of people with diabetes in Portugal are presented in Table 2.

Table 2. Prevalence and estimated number of people with diabetes (20-79 years) for 2013 and 2035 in Portugal (adapted from (5)).

Prevalence adjusted to the national population (%)		**Prevalence adjusted to the world population (%)**		**Diabetes cases (20-79 years) in 1000s**		**Mean annual increment (000s)**	**Proportional change in number of people with diabetes from 2013 to 2035 (%)**
2013	**2035**	**2013**	**2035**	**2013**	**2035**		
13.0	15.8	9.6	9.8	1032	1233	9	19.5

This increase in prevalence can be due to a variety of factors that influence the risk of developing T2D (6). As a multifactorial disease, T2D is caused by the interaction of genetic susceptibility and environmental factors causing hyperglycemia, the hallmark of T2D (Figure 1) (6,7). The main environmental factors that contribute to the development of this disorder are a sedentary lifestyle and increased caloric intake that are responsible for most of the weight excess and obesity (8).

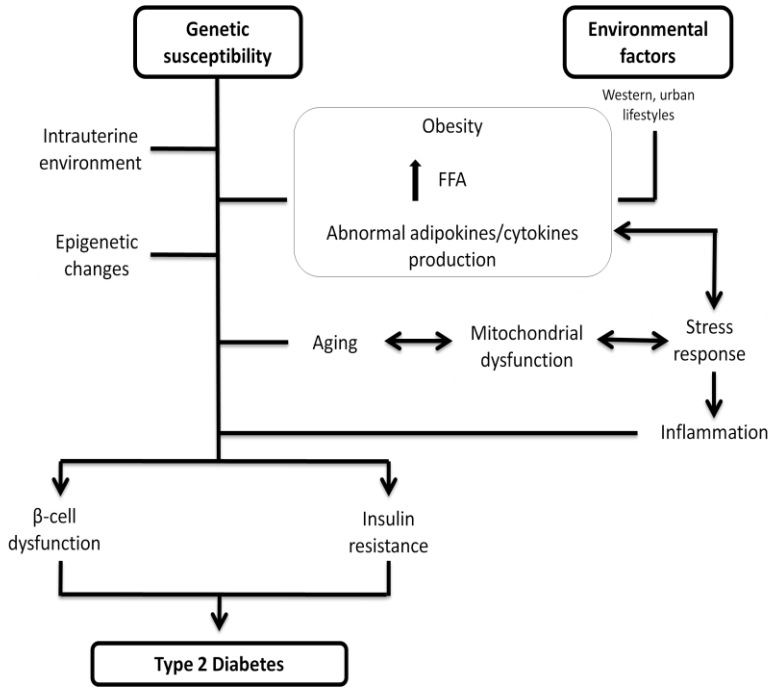

Figure 1. Overview of the interaction between the risk factors underlying the development of T2D (adapted from (9)). FFA: Free fatty acids.

2. Pathophysiology of Type 2 Diabetes

From a pathophysiological standpoint, T2D is characterized by increased levels of blood glucose due to impairment in insulin action and/or insulin secretion (10). Insulin, the main hormone that regulates glucose uptake from the blood into most cells, is synthesized in the pancreatic β-cells (1). The pancreas is an organ constituted by about 1,000,000 islets of Langerhans, each with approximately 1,000 β-cells. These islets are highly organized aggregates of β-cells (production and release of insulin), α-cells (production and release of glucagon), δ-cells (production and release of somatostatin) and pancreatic polypeptide (PP) cells (production and release of the pancreatic polypeptide) (11,12,13). The islets receive a rich systemic vascular supply (approximately 10% of the blood supply) through the splenic and superior mesenteric arteries, which allows the β-cells to sense changes in plasma glucose concentration and release insulin to match the metabolic demand (12,14).

Therefore, the development of T2D is characterized by a combination of insulin resistance and β-cell dysfunction, which result from a complex interaction between predisposing genetic factors and precipitating environmental ones (9,15).

2.1. Genetic and environmental factors

The role of the genetic component in T2D is well documented (9). The risk of developing this disorder increases when relatives have the disease; 15% to 25% of first-degree relatives of T2D patients develop IGT or diabetes (16). Moreover, there is also a 40% chance of developing T2D for those who have an affected parent (higher if it is the mother rather than the father) and 70% chance if both parents are diabetics (17). In addition, a genetic predisposition to T2D is also supported by the observation that differences in disease prevalence rates exist among populations, even after migration of entire ethnic groups to another country, and thus independent from the environmental influences. On the other hand, the role of the environmental factors in T2D susceptibility is equally well known, being the spread of the western way of life in developing countries an explanation for the disease epidemic explosion (9).

Over the past few years various methods, such as linkage analysis, candidate gene association, genotyping for genetic markers, meta-analysis and genome-wide association studies (GWAS), were used to highlight the role of genetic factors in the development of T2D (9). More than 60 genetic variants, more specifically single nucleotide polymorphisms (SNPs), have been associated with this disorder (9,18,19). A list of the susceptibility genes associated with T2D is available in Appendix A – Table A1. However, some of these variants are non-coding, thus becoming a challenge to analyze their functional impact. It is estimated that the genetic variants identified so far only explain about 10% of T2D heritability (9,19,20).

Most of the variants identified so far are directly related to insulin secretion (β-cell function) and not insulin action in tissues sensitive to insulin, suggesting that most of the risk associated with T2D relates to genetic defects in pancreatic β-cells (9,20). One of the mechanisms responsible for β-cell dysfunction and resulting insulin secretion deficiency is the exposure to an adverse intrauterine environment and consequent low birth weight. This adversity affects fetal development by modifying gene expression of both pluripotent cells, that are rapidly replicating, and terminally differentiated cells, which replicate poorly. However, whether the effects of exposure to an altered intrauterine milieu extend into adulthood depends on whether the cells are undergoing differentiation, proliferation and/or functional maturation at the time of that exposure. There are several examples where human exposure to an abnormal intrauterine milieu leads to abnormalities in glucose

homeostasis and ultimately T2D. For example, an epidemiological study from Hertfordshire, in the United Kingdom, found that men who were the smallest at birth (< 2.5 kilograms (kg)) were seven times more likely to have glucose intolerance or T2D than those who were heaviest at birth. Moreover, a study demonstrated that perinatal nutrient restriction leads to silencing histone modifications in the skeletal muscle, which in turn directly contributes to a decrease in the expression of the *SLC2A4* gene. This gene encodes the glucose transporter type 4 (GLUT4) and a decrease in its expression creates a metabolic knockdown of this important regulator of peripheral glucose transport and insulin resistance, contributing to the adult T2D phenotype. Therefore, maternal nutrition can change the stable expression of genes in the offspring potentially through epigenetic modifications occurring in utero. Those modifications of the genome provide a mechanism that allows the stable propagation of gene expression from one generation of cells to the next, through histone modifications and deoxyribonucleic acid (DNA) methylation (21).

The nucleosome is composed of DNA wrapped around an octameric complex that consists of two molecules of each one of the four histones: H2A, H2B, H3 and H4. The N-termini of histones can be modified by several processes, being the most common ones acetylation and methylation of lysine residues in this terminal of H3 and H4. Increased acetylation induces transcription activation, whereas decreased acetylation is usually associated with transcription repression. On the other hand, methylation of histones is related to both transcription repression and activation. The other mechanism through which epigenetic information is inherited is DNA methylation (21). Methylation is a key epigenetic feature of DNA that plays an important role in chromosomal integrity and the regulation of gene expression with different methylation profiles now being associated with several complex diseases (22). In this mechanism, a cytosine base is modified by DNA methyltransferase at its C5 position. In addition to targeted DNA methylation changes in response to external stimuli, random DNA methylation changes also occur during aging in several tissues and are associated with increased oxidative stress. Such changes in the DNA methylation patterns can affect the expression of multiple genes. An example of that is the methylation within the promoter of the *PDX1* gene, a pancreatic homeobox transcription factor, induced by overexpression of a DNA methyltransferase. This methylation resulted in an effected gene expression, demonstrating that genes

essential to the pancreatic β-cell development, such as the *PDX1* gene, are susceptible to epigenetic modifications (21).

Exposure to oxidative stress can directly mediate both DNA methylation and chromatin remodeling in multiple diseases and thus could be suggested as a mechanism by which aberrant epigenetic programming leads to T2D (21). Furthermore, there is growing evidence that suggests that DNA methylation differences are associated with complex diseases and that understanding the role of developmental programming of genes crucial to the development of T2D might unveil a critical window during which epigenetic therapeutic agents could be used as a means to prevent the later development of the disease. Therefore, epigenetic modification provides a dynamic link between each individual's genetic background and relevant environmental exposures, ultimately affecting the expression of key genes linked to the development of T2D, including genes critical for pancreatic development, β-cell function, peripheral glucose uptake and insulin resistance (21,22).

Genetic variants present in other genes were also associated with impaired insulin secretion. For instance, in the *TCF7L2* gene, a transcription factor, and in the *CDKAL1* gene, a cyclin-dependent kinase 5 (CDK5) regulatory subunit, genetic variants were associated with reduced fasting insulin, which indicates β-cell dysfunction (23). Other genes associated with increased T2D risk are *KCNJ11* gene, which protein products take place in the formation of adenosine triphosphate (ATP)-sensitive potassium channel/sulfonylurea receptor of the pancreatic β-cells and the *SLC2A2* gene. This latter gene encodes the glucose transporter type 2 (GLUT2), which is expressed in the pancreatic β-cells, liver and kidney, and functions as a glucose sensor to maintain glucose homeostasis. Furthermore, a polymorphism in the gene that encodes the insulin receptor substrate 1 (IRS-1) was also associated with T2D. This polymorphism causes a defect in the binding of the p85 subunit of the phosphatidylinositol 3-kinase (PI3K), which leads to a decrease in insulin secretion in response to glucose and sulfonylureas (9).

The mechanisms responsible for impaired insulin action in tissues sensitive to insulin include obesity; the secretion of circulating factors by adipocytes, particularly of free fatty acids (FFA); mitochondrial dysfunction as a consequence of oxidative stress and inflammation (9). The modern lifestyle has been a major contributor for the increased prevalence of obesity. This factor, combined with advancing age, forms the most potent

risk factors for T2D (24). An increased caloric intake and a sedentary lifestyle, two conditions common in populations with a higher standard of living and a more westernized lifestyle, are responsible for most of the excess weight and obesity in the modern adult's life (9). Obesity is a state of low grade inflammation and constitutes a key risk factor for T2D, as it desensitizes glucose recipient organs to the action of insulin (25). Therefore, obesity predisposes to insulin resistance, which represents the initial step in the natural history of T2D (9). For example, it has been calculated that a lean, insulin sensitive adult may need as little as 0.5 units of insulin to dispose of an oral load of 75 g of glucose over 2 hours, while an obese, insulin resistant, glucose intolerant subject may require 45 units to perform the same task (11). Genetic variants in the gene associated with obesity, the *FTO* gene, appear to influence predisposition to T2D through a positive effect on body mass index and obesity (9). This gene is associated with higher fasting insulin, which is consistent with a primary defect in insulin action (9,23). Moreover, also the *KLF14* and *PPARG* genes appear to have primary effects in insulin action, but, unlike the *FTO* gene, those effects are not driven by obesity (9).

The insulin-stimulated glucose uptake in adipocytes depends mostly on the transporter GLUT4. The downregulation of this transporter expression in adipose tissue can be found in obesity and T2D, and is commonly associated with insulin resistance (Figure 2) (10).

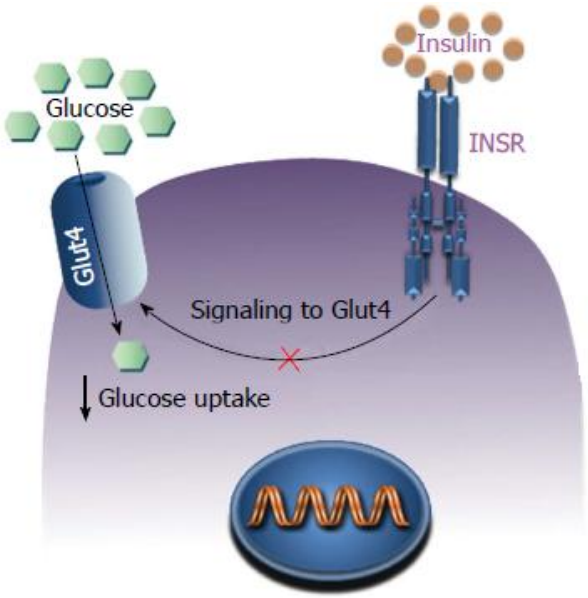

Figure 2. Effects of the GLUT4 downregulation in adipose tissue (adapted from (9)). GLUT4: glucose transporter type 4; INSR: insulin receptor.

However, given that the skeletal muscle is the major site for glucose disposal, the development of hyperglycemia associated with obesity and T2D, cannot be only explained by the decreased glucose uptake in the adipose tissue due to the downregulation of GLUT4 in adipocytes. Therefore, the secretion of circulating factors that mediate cross-organ communication ("crosstalk") by adipocytes is suggested for explaining T2D pathogenesis. Those factors, such as adiponectin, resistin, leptin, visfatin, retinol-binding protein 4 (RBP4) and the cytokines, interleukin-6 (IL-6) and tumor necrosis factor α (TNFα), may alter insulin action (10). Furthermore, the reduced glucose uptake in adipocytes accelerates lipolysis and raises the level of FFA, which also contributes to the altered insulin action (10,26).

Over the past few years, an increasing number of studies have linked lipid accumulation in the skeletal muscle to reduced insulin sensitivity in type 2 diabetic individuals (27). The disparity between the uptake of fatty acids and their oxidation induce the accumulation of triacylglycerol and fatty acid metabolites such as fatty acyl-coenzyme A (acyl-CoA), diacylglycerol and ceramide in the skeletal muscle. This accumulation leads to an elevation in circulating FFA, which inhibit insulin action via activation of serine/threonine kinases and phosphorylation of serine residues of the IRS-1 (1,27). The phosphorylation reduces tyrosine phosphorylation of IRS-1 in response to insulin, and thus signaling downstream of IRS-1 (10). Therefore, the increase of FFA is associated with a decrease in insulin signaling and glucose disposal rates (1). In addition to FFA, also the TNFα secreted by adipocytes, stimulates phosphorylation of the serine residues of IRS-1 (10). Furthermore, there is growing evidence that the oxidative capacity of the skeletal muscle, which is mostly dependent on mitochondrial function, is directly correlated with insulin sensitivity, being the impaired ability to oxidize fatty acids associated with insulin resistance. A reduction in the number, as well as changes in the morphology of mitochondria, has been observed in the skeletal muscle of type 2 diabetic patients. This suggests that excessive FFA could alter mitochondrial functions, leading to an increased production of reactive oxygen species (ROS), and consequently oxidative stress. Therefore, mitochondrial dysfunction is a consequence of the ROS production induced by hyperglycemia and increased FFA in the skeletal muscle. The full effects of that FFA excess on mitochondrial biogenesis and functions have not been investigated in detail. Moreover, because increased ROS levels can also play an important role in altered insulin secretion by the pancreas,

oxidative stress contributes to both insulin resistance as well as pancreatic β-cell dysfunction (27). The role of the recently identified T2D-associated *SREBF1* gene in the susceptibility for this disorder is still unclear, however, it is known that this gene is involved in the transcriptional regulation of lipid homeostasis (9).

The concept that obesity and elevated cytokine production results in an inflammatory state that can cause of insulin resistance has emerged. Inflammatory pathways can be initiated by both extracellular mediators or by intracellular stresses. The extracellular mediators include cytokines and lipids while the intracellular stresses can be endoplasmic reticulum (ER) stress or excess ROS production by mitochondria. The increased glucose metabolism can lead to a rise in the mitochondrial production of ROS. This production is also elevated in obesity (10). Furthermore, the adipose tissue is a major source of inflammation with the infiltration of macrophages as the primary source of cytokines in obese individuals (25). This contributes to an enhanced activation of the inflammatory pathways (10). The obesity-induced macrophages infiltration is represented in Figure 3.

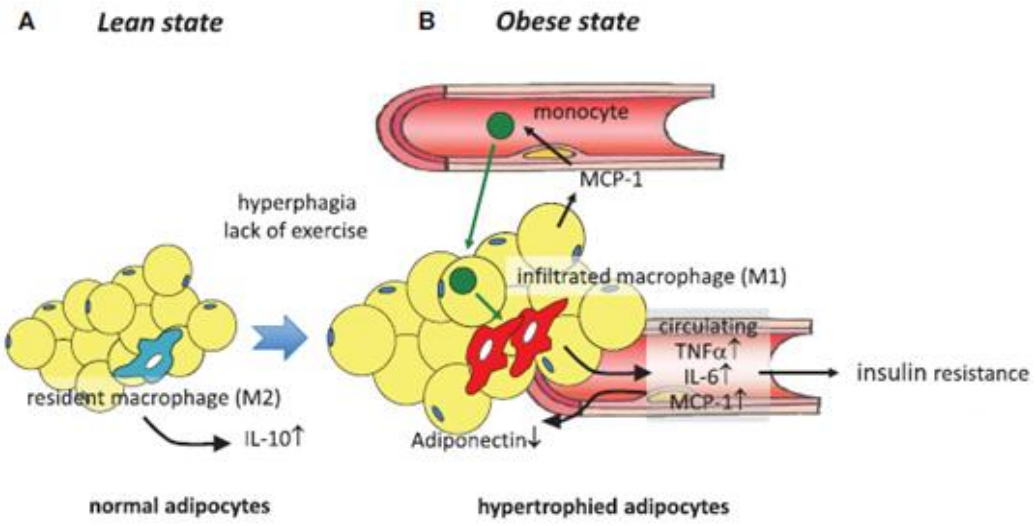

Figure 3. Macrophage infiltration into adipose tissue induced by obesity (adapted from (28)). (A) adipose tissue in a lean state, most resident macrophages are M2 macrophages that contribute to insulin sensitivity by secreting IL-10, (B) hyperphagia and lack of exercise cause hypertrophy of adipocytes, which induces the secretion of MCP-1 to the circulation, leading to the recruitment of circulating monocytes to adipose tissues. Those monocytes differentiate into activated M1 macrophages and secrete proinflammatory cytokines such as TNFα, IL-6 and MCP-1, therefore, contributing to inflammation in adipose tissue and a decrease in the adiponectin levels. That results in insulin resistance. IL-6: interleukin-6; IL-10: interleukin-10; MCP-1: monocyte chemotactic protein-1; TNFα: tumor necrosis factor α.

Therefore, both the adipose tissue and the macrophages within that tissue serve in both endocrine and paracrine fashion to promote inflammation and decrease insulin sensitivity. The inhibition of signaling downstream of IRS-1 is thought to be the primary mechanism through which this inflammatory state causes insulin resistance (10).

A better understanding of the genetics in the T2D pathophysiology will allow the translation of the genetic information to the clinical practice, and the identification of new opportunities for treatment, diagnosis and monitoring. However, the information available so far does not provide sufficient evidence to support the use of genetic screening for the prediction of T2D (18,20).

2.2. Insulin resistance and β-cell dysfunction

Functionally, β-cells are the most important endocrine cells, having as the main product the hormone insulin (11). The biosynthesis of this hormone is controlled by multiple factors, being glucose metabolism the main one (14). Pancreatic β-cells are electrically excitable and use changes in the membrane potential to stimulate or inhibit insulin secretion (12). The objective of this β-cell electrical activity is to elevate the intracellular concentration of calcium ($[Ca^{2+}]i$) promoting exocytosis of insulin granules by merging to the plasma membranes and releasing insulin (12,14). After a meal, elevated glucose levels trigger a signaling cascade in the pancreatic islet β-cells to release insulin. This process was described as an 8 step process (Figure 4) (29).

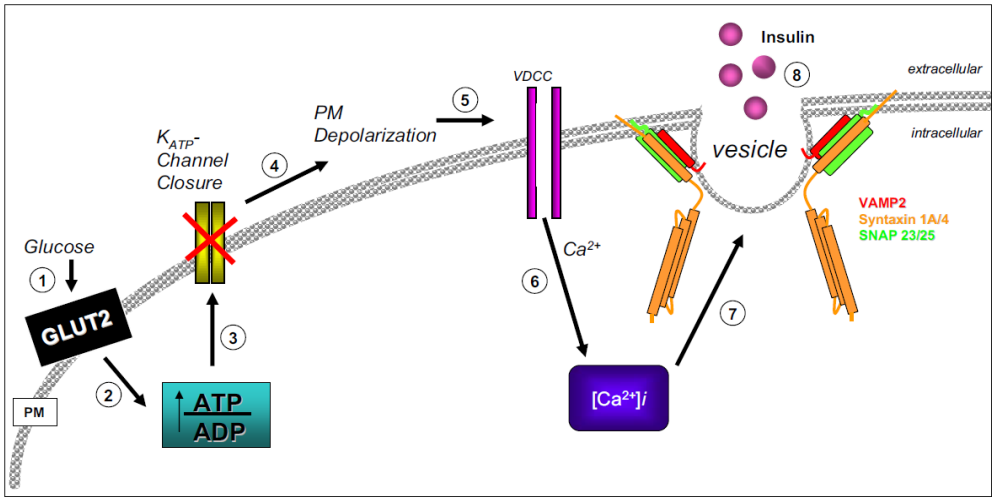

Figure 4. Glucose stimulated insulin release from the pancreatic β-cells (taken from (29)). ADP: adenosine diphosphate; ATP: adenosine triphosphate; Ca^{2+}: calcium ion; $[Ca^{2+}]i$: intracellular calcium concentration; GLUT2: glucose transporter type 2; K_{ATP}: adenosine triphosphate-dependent potassium channels; PM: plasma membrane; SNAP 23/25: synaptosomal-associated protein 23/25 kilodalton; VAMP2: vesicle-associated membrane protein 2; VDCC: voltage-dependent calcium channels.

Glucose-stimulated insulin secretion from the β-cells is mediated by the soluble N-ethylmaleimide-sensitive factor activating protein receptor (SNARE) complex proteins, syntaxin 1 and 4, synaptosomal-associated protein 23 kilodaltons (kDa) (SNAP23) and synaptosomal-associated protein 25 kDa (SNAP25) and lastly, vesicle-associated membrane protein 2 (VAMP2). In step 1, glucose entries β-cells through the constitutively active plasma membrane (PM) localized GLUT2 transporter. Following, the glucose suffers phosphorylation by glucokinase to generate glucose-6-phosphate. This latter is subsequently metabolized via mitochondrial oxidative phosphorylation. This metabolization increases the intracellular ATP: adenosine diphosphate (ADP) ratio (step 2). Elevated β-cell ATP levels induce closure of the ATP-dependent potassium channels (K_{ATP}) (step 3). This closure results in PM depolarization (step 4) and an influx of calcium ions (Ca^{2+}) through the voltage-dependent calcium channels (VDCC) (step 5) to yield a raise in $[Ca^{2+}]i$ (step 6). The increase in [Ca2+]*i* signals SNARE complex formation (step 7) which, through a largely uncharacterized series of events, mediate vesicle fusion to facilitate insulin release from the granules (step 8). The precise mechanism by which Ca^{2+} triggers granule fusion remains unclear (29).

In a normal situation, insulin secretion from the β-cells occurs in a pulsatile fashion in synchronization with Ca^{2+} influxes (29). This secretion possesses a biphasic pattern with a momentary first phase (about 10 minutes) and a sustained second one (14). The first phase occurs within 5-10 minutes following β-cell stimulation, which happens after a spike in blood glucose concentration (for example after a meal). These cells are stimulated to release insulin in large bursts (ultradian oscillations) to maximize the efficiency of glucose clearance to the target tissues and suppress the endogenous glucose production, which occurs primarily in the liver (15,29,30). In the second phase, insulin is released every 11 to 14 minutes to keep a normal regulation of endogenous glucose production (31). This phase is less robust than the first phase, but can be sustained for several hours if elevated blood glucose levels persist. These phases of secretion are thought to use separate pools of insulin-containing granules. The insulin secreted in the first phase appears to arise from PM pre-docked granules, termed the readily releasable pool (RRP), while the insulin secreted in the second phase is believed to involve release from a granule pool deeper within the cell, the storage-granule pool. This latter pool presumably replenishes the RRP. In addition to this, the insulin release phases also differ in their required SNARE complex proteins. The first phase insulin release uses syntaxin 1A, syntaxin 4, SNAP25 or SNAP23, and VAMP2, whereas secretion in the second phase involves syntaxin 4, SNAP25 or SNAP23, and VAMP2, but specifically not syntaxin 1A (29). Usually after the blood glucose level lowers, the secretion of insulin also diminishes, and glucagon is released to stimulate the conversion of the stored glycogen into glucose. The release of glucagon usually happens after a fasting period (1).

The plasma insulin concentrations are dependent on insulin secretion from the pancreatic β-cells and hepatic and renal clearance, and can be influenced by several factors such as size and composition of meals and changes in insulin sensitivity of the target tissue. Whole-body insulin clearance occurs mostly through hepatic degradation since the kidneys only contribute with about 20% of the clearance. This process is regulated by the insulin concentration through a saturable process (11).

In general, insulin resistance is defined as impaired insulin-mediated glucose clearance into target tissues, especially the skeletal muscle (15). As insulin resistance emerges, there is a failure of adequate insulin-mediated glucose clearance from the skeletal muscle combined with a reduction in the suppression of hepatic glucose production, which leads to

hyperglycemia (32). Besides this, renal and hepatic failures lead to a reduced organ clearance of insulin, also contributing to the hyperglycemia caused by insulin resistance. Furthermore, this hyperglycemia is also responsible for an increase in insulin secretion from the β-cells as a compensatory response, which ultimately causes hyperinsulinemia (11,32). Hyperinsulinemia occurs at an early stage of pancreatic β-cell dysfunction, followed by β-cell failure (33).

The main theories associated with β-cell failure involve the exposure of β-cells to glucose (glucotoxicity) and FFA (lipotoxicity). Although the mechanisms are not fully understood, it is known that chronic exposure to glucose leads to an increase in cytosolic Ca^{2+} that induce β-cell destruction and a long exposure of the islets to FFA leads to a decrease in insulin secretion. Moreover, the IRS-1 and insulin receptor substrate 2 (IRS-2) are important to β-cell function and survival, being the insulin receptor signaling cascade crucial for regulation of islets cell differentiation and function (14,34).

One of the major defects that can be observed in β-cell failure and that impairs insulin secretion is the disruption of the pulsatile patterns of insulin release in the early course of the disease (31). This, in combination with the loss of insulin secretion burst after an acute rise in glucose concentration, causes glucose intolerance (32). Defects in proinsulin processing at the β-cell/islet level were also found, although the underlying mechanism is yet unidentified (35). Another change that some but not all studies report is the decrease in β-cell relative mass. The underlying mechanism for this reduction has been suggested as increased β-cell apoptosis since it is the only mechanism altered and both new islet formation and β-cell replication are normal in individuals with T2D. It has been found that there is a 35% to 39% decrease in β-cell mass in T2D patients compared to non-diabetic individuals (35,36). Although there are many proposed mechanisms for β-cell dysfunction, none of them is widely accepted. What is already known is that β-cell dysfunction causes a rise in the blood glucose levels, leading to impaired glucose tolerance and/or impaired fasting glucose, and consequently T2D (37).

The relationship between insulin resistance and β-cell dysfunction is relevant due to the feedback that links them both. Insulin resistance signals back to the pancreatic β-cells to increase the insulin output to maintain normal glucose tolerance (11). Currently, it is considered that T2D is a dual defect disease (38).

3. Complications associated with Type 2 Diabetes

Due to the slow onset of symptoms in T2D, the condition can remain undetected for several years. During that time, increased levels of untreated blood glucose can cause the development of irreversible complications (39). These complications appear due to injuries in the vasculature, being classified as either microvascular (nephropathy, retinopathy and neuropathy) or macrovascular (development of atherosclerosis that affects vital organs such as the heart and brain, causing cardiovascular and cerebrovascular disease) (39,40,41). Both injuries cause elevated disability and healthcare costs and are responsible for most of the morbidity and mortality associated with diabetes (24).

Currently, even though it is known that hyperglycemia induces damage to the particular cell subtypes, such as the mesangial cells in the renal glomerulus, the capillary endothelial cells in the retina and the neurons and Schwann cells in the peripheral nerves, little is known about the mechanisms underlying these damages (24,40). The activation of key metabolic and hemodynamic pathways by hyperglycemia can lead to the development of diabetic complications since each of those pathways can generate toxic and reactive metabolites that can then alter several key signaling intracellular pathways. Thus, several hypotheses for the responsible mechanisms have been proposed (summarized in Figure 5) (40).

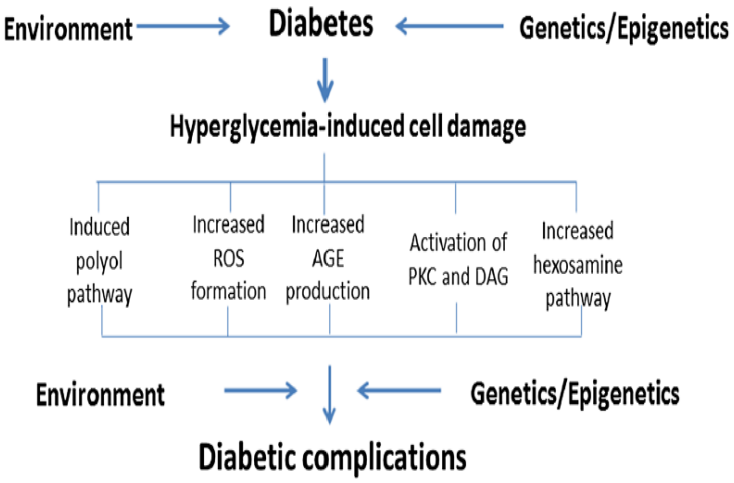

Figure 5. Overview of the metabolic pathways involved in the pathogenesis of complications associated with T2D (adapted from (40)). AGE: advanced glycation end-product; DAG: diacylglycerol; PKC: protein kinase C; ROS: reactive oxygen species.

The main hypothesis considers the increased flux to the aldose reductase (AR) or polyol pathways due to elevated intracellular glucose concentration. The activation of AR (highly expressed in the lens compared to other tissues) and polyol pathways increase nicotinamide adenine dinucleotide phosphate (NADPH) consumption and decrease the production of glutathione (GSH). This latter is an important antioxidant mechanism in the cell, leaving the cells with an increased susceptibility to oxidative stress (40). Other mechanisms involved are the increased advanced glycation end-products (AGEs) formation; the overproduction of ROS by both enzymatic and non-enzymatic processes, being its major source the mitochondria and the NADPH oxidase complex; the increased hexosamine pathway flux and activation of protein kinase C (PKC) pathway (40,42). The increase in oxidative stress that leads to the overproduction of ROS also increases superoxide and reduce nitric oxide levels. The excess of superoxide during oxidative stress partially cause an increase in diacylglycerol (DAG), an activator of the PKC pathway, and methylglyoxal (MG), the main intracellular AGE precursor. The activation of the PKC pathway causes changes in the blood flow and increases the vascular permeability, ultimately causing vascular damage in the kidney, nerves and retina (40).

4. Diabetic Nephropathy

Approximately 25% to 30% of patients with T2D develop diabetic nephropathy, being this prevalence predicted to increase in the decades ahead (43,44). The kidneys regulate electrolyte, water and acid-base balance, being, therefore, indispensable for the maintenance of body homeostasis. These functions are carried out by the collective action of approximately 1,000,000 nephrons in each kidney, each consisting of a glomerulus and a system of renal tubules. The glomerulus is responsible for the filtration process while the tubular system regulates selective reabsorption and secretion, dictating the final composition of urine. Normally, the kidneys ensure almost protein-free and ultra-filtrated urine, however, in disease conditions, a variety of plasma proteins get excreted in the urine (45). The structure of the nephron is represented in Figure 6.

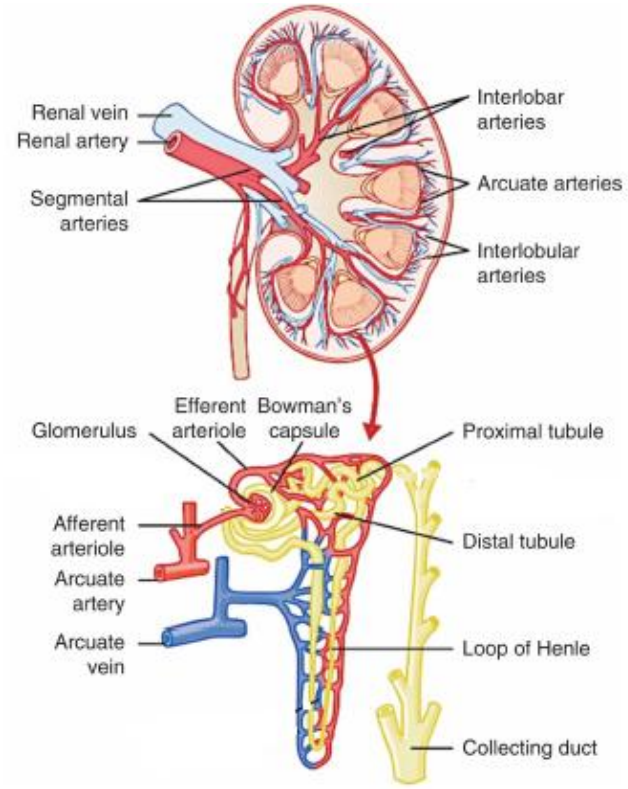

Figure 6. Kidney nephron structure (adapted from (46)).

Diabetic nephropathy is defined by a progressive increase in proteinuria (> 500 milligrams (mg)/24 hours), the leakage of valuable plasma proteins, mainly albumin, into the urine, and a gradual decline in renal function, being the leading cause of chronic kidney disease and end-stage renal disease (ESRD) (43,47,48,49). Diabetic nephropathy has been categorized into two different phases, microalbuminuria and macroalbuminuria, based on the values of urinary albumin excretion (UAE) (47). The UAE values are also used to make the clinical diagnosis of diabetic nephropathy instead of an invasive renal biopsy (50). When a patient has microalbuminuria (30-299 mg/24 hours), it is a sign of early diabetic nephropathy, also called incipient diabetic nephropathy. If not treated, it can evolve to macroalbuminuria (> 300 mg/24 hours), also known as overt or clinical diabetic nephropathy and a sign of the complication progression (51). Besides the renal damage caused by this microvascular complication, diabetic nephropathy can also be associated with increased cardiovascular mortality. Most of the patients die due to cardiovascular causes before even progressing to renal failure (52,53).

4.1. Risk factors and disease pathophysiology

The main risk factors that lead to the initiation and progression of nephropathy are hyperglycemia, dyslipidemia (54), systemic hypertension (55) and dietary factors such as the amount and source of protein (56). Besides this, other microvascular complications, such as diabetic retinopathy, can be a predictor for the development of diabetic nephropathy since both of these complications share common risk factors like the lack of a proper glycemic control, elevated blood pressure and dyslipidemia (54). Furthermore, since about 7% of T2D patients already present microalbuminuria at the time of disease diagnosis (47) and due to the possibility of patients with microalbuminuria developing macroalbuminuria, which can ultimately lead to ESRD, microalbuminuria can also be considered a risk factor for diabetic nephropathy progression (51,57). These risk factors only explain 30% to 50% of the variation in the progression of this complication, being thought that other factors such as genetic factors may contribute to determining susceptibility to diabetic nephropathy (40,50). However, it is still difficult to clearly define a genetic component in diabetic nephropathy, since the studies performed so far have not been successful in identifying genes that consistently show an association with the disease (49,58).

The natural history of diabetic nephropathy is characterized by a prolonged period of clinical silence during which subtle renal changes emerge. For this reason, when the clinical diagnosis of the complication is made, significant kidney damage (structural and/or functional) has already developed (50,59). Structurally, kidney abnormalities include hypertrophy of the kidney, tubular atrophy, and, more importantly, extracellular matrix (ECM) accumulation and dysfunction and/or loss of podocytes, the two main changes that occur in diabetic nephropathy (44,48,60,61,62). Regarding functional modifications, there is an increase in the glomerular filtration rate (GFR) (glomerular hyperfiltration) along with a raise in the glomerular capillary hydraulic pressure (P_{GC}), followed by systemic hypertension, a gradual decline in the GFR over the period of several years that only manifests when structural renal lesions become far advanced, and eventual loss of renal function (59,62,63,64). The different factors that play a role in the pathophysiology of diabetic nephropathy and lead to renal damage are presented in Figure 7.

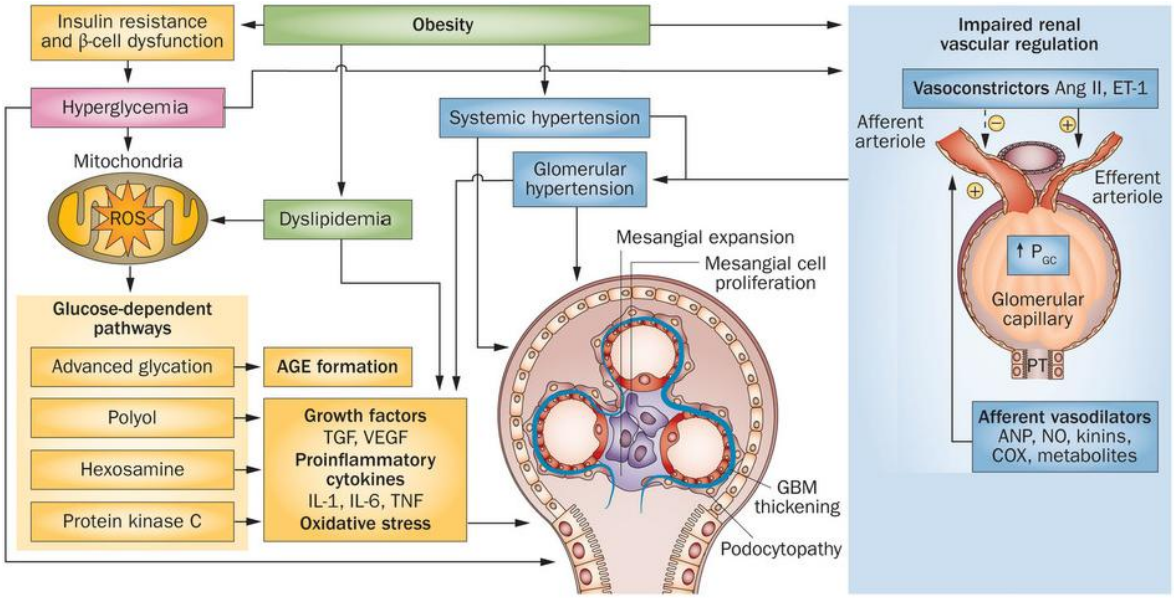

Figure 7. Overview of the hemodynamic and metabolic factors that influence the pathophysiology of diabetic nephropathy (adapted from (65)). Obesity and hyperglycemia alter vasoactive regulators of the afferent and efferent arterioles, leading to increased P_{GC}, hyperperfusion, and hyperfiltration. These early renal hemodynamic changes, combined with systemic hypertension, result in glomerular damage, which contributes to the development and progression of diabetic nephropathy. Additionally, chronic hyperglycemia and dyslipidemia induce mitochondrial superoxide production, and consequently elevation of ROS, leading to oxidative stress. This activates several pathways that also lead to the development and progression of diabetic nephropathy. Collectively, these factors in the diabetic milieu are responsible for inducing glomerular damage, histologically characterized by mesangial expansion, thickening of the glomerular basement membrane and podocytopathy. AGE: advanced glycation end-product; Ang II: angiotensin II; ANP: atrial natriuretic peptide; COX: cyclooxygenase; ET-1: endothelin-1; GBM: glomerular basement membrane; IL-1: interleukin-1; IL-6: interleukin-6; NO: nitric oxide; P_{GC}: glomerular capillary hydraulic pressure; ROS: reactive oxygen species; TGF: transforming growth factor; TNF: tumor necrosis factor; VEGF: vascular endothelial growth factor.

The development of these renal abnormalities in diabetic nephropathy was categorized into a series of stages (Table 3), however, most of the time, it is difficult to document those different stages in a diabetic patient due to the use of confounding factors, such as systemic hypertension medication, which ultimately alters the natural course of diabetic nephropathy (62,66).

Table 3. Stages of diabetic nephropathy in type 2 diabetic individuals (adapted from (62,66)).

Stage	Stage characteristics	Chronology
1 – Early hyperfunction and hypertrophy	Glomerular hyperfiltration	Present at the time of diagnosis
2 – Silent Stage	Thickening of the glomerular basement membrane and mesangial matrix expansion	First 5 years
3 – Incipient diabetic nephropathy	Microalbuminuria	6-15 years
4 – Overt diabetic nephropathy	Macroalbuminuria	15-25 years
5 – Uremic	ESRD	25-30 years

ESRD: End-stage renal disease.

These changes, both structural and functional, can be detected by direct or indirect methods. Directly, alterations in the kidney can be found by histological alterations in the renal biopsy, whether indirectly, changes can be discovered by a rise in UAE, modifications in the urine sediment or differences identified with imaging techniques (43).

4.1.1. Extracellular matrix accumulation

The mesangial cells occupy a central position in the renal glomerulus, forming the glomerulus central stalk (67,68). These cells generate and are embed in their own ECM, known only as the mesangial matrix. Along with their matrix, mesangial cells are part of a functional unit interacting closely with endothelial cells and podocytes, meaning that alterations in one type of cell can produce changes in the others (68). The mesangial cells resemble vascular smooth muscle cells in phenotype and responsiveness to different stimuli, being responsible for providing structural support for the glomerular capillary loops (68,69). Due to that smooth muscle similarity, these cells possess contractile properties (68). Therefore, they can play a role in glomerular contraction, having the ability to contract or relax in response to some vasoactive agents. That ability allows them to modify glomerular filtration locally, through modulation of the ultrafiltration coefficient and capillary filtration surface area, thus helping to regulate GFR (67,69). The filtration

surface area can be regulated by a constriction of the capillary loops that results in the reduction of the capillary diameter and consequently alters the intraglomerular capillary flow (68,69). Besides the functions associated with its smooth muscle cell characteristics, mesangial cells are also capable of some other functions (67). They can generate and control the turnover of the mesangial matrix and serve as a source and target of growth factors. Therefore, mesangial cells and their matrix can be involved in the pathophysiology of some glomerular diseases (68).

The composition and amount of mesangial matrix are tightly controlled in health but can be markedly altered during disease. An abnormal production of the cytokines regulating the various matrix components emanating from mesangial cells, particularly transforming growth factor β (TGF-β), are responsible for the mesangial matrix expansion as well as glomerular basement membrane thickening that are associated with diabetic nephropathy (68,69). The glomerular basement membrane changes are a result of the interplay between metabolic and hemodynamic factors in the renal microcirculation. As a rule, the thickening of the glomerular basement membrane occurs soon after the onset of T2D and gradually increases as diabetic nephropathy clinically manifests (70). The structure of a normal glomerulus is presented in Figure 8.

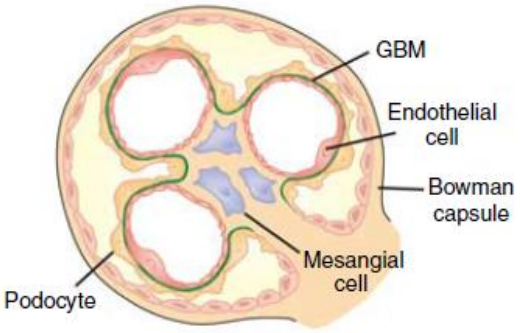

Figure 8. Structure of the renal glomerulus (adapted from (71)). GBM: glomerular basement membrane.

Extracellular matrix remodeling depends on the balance between the synthesis of ECM proteins and their degradation, a process controlled by a complex cytokine network in which the TGF-β has a major role (72). The main process during ECM remodeling is cleavage of the matrix components, which is important for regulating ECM abundance, composition and structure, as well as for releasing biologically active molecules, such as

growth factors. The ECM can be cleaved by different families of proteases, being matrix metalloproteinases (MMPs) the main ones. Most MMPs are secreted as zymogens and are subsequently activated in the extracellular space. Their activation primarily occurs via proteolytic cleavage, by serine proteases or other MMPs, or via the modification of the thiol group by oxidation, for example through ROS. Furthermore, they are produced as either soluble or cell membrane-anchored proteases and cleave ECM components with wide substrate specificities (73). On the basis of substrate specificity, sequence homology and domain organization, MMPs have been classified into six groups: collagenases (MMP1, MMP8, MMP13 and MMP18), gelatinases (MMP2 and MMP9), stromelysins (MMP3, MMP10 and MMP11), matrilysins (MMP7 and MMP26), membrane-type MMPs (MT-MMPs) (MMP14, MMP15, MMP16, MMP17, MMP24 and MMP25) and other MMPs (MMP12, MMP19, MMP20, MMP21, MMP23, MMP27 and MMP28), which are not classified in any of the other categories. Moreover, MT-MMPs can be divided into type I transmembrane proteins and glycosylphosphatidylinositol (GPI)-anchored proteins (74). Collectively, the MMPs can degrade all the ECM proteins (73).

Under normal physiological conditions, the activity of MMPs is low, but tightly regulated at the level of transcription, activation of the precursor zymogens, interaction with specific ECM components and inhibition by endogenous inhibitors (73,74). During repair or remodeling processes, this activity increases. The ECM proteolysis also requires tight regulation to avoid excessive tissue degradation, being the activity of endogenous inhibitors that inactivate MMPs important for tissue integrity. The tissue inhibitor of metalloproteinases (TIMPs) family consists of four members (TIMP1–TIMP4) that reversibly inhibit the activity of MMPs. The ratios MMP:TIMP determine the overall proteolytic activity, and each TIMP displays preferential MMP-binding specificity (73). Therefore, both increased, as well as decreased MMP activity may result in dysregulated ECM remodeling that leads to a variety of diseases, such as cancer, atherosclerosis, nephritis, tissue ulcers and fibrosis (73,74).

In a hyperglycemic state, MMP gene expression can be regulated via various transcription factors, such as activator protein-1 (AP-1) and nuclear factor kappa B (NF-kB), or depending upon the tissue levels of growth factor cytokines, such as TGF-β and connective tissue growth factor (CTGF). In the past years, extensive investigatory efforts have been made to understand the role of MMPs in a variety of kidney diseases, including

diabetic nephropathy. However, their role in kidney pathology seems to be still very much confounding (75).

In mammalians, TGF-β has three different isoforms, TGF-β1, TGF-β2 and TGF-β3, of which TGF-β1 is the most potent promoter of ECM accumulation (44). This isoform is secreted as latent complexes, which are then stored in the ECM to provide stability to its active form and a constant activable source (76).

In the kidney, the mesangial cells secret TGF-β1 in a latent dimeric complex form, containing in the C-terminal the mature TGF-β and in its N-terminal the inactive domain TGF-β latency-associated peptide (LAP) (77,78). The two polypeptide chains of inactive TGF-β associate to form a disulfide-bonded dimer and LAP remains associated with that dimer by non-covalent interactions, which results in the formation of what is known as the small latent complex. In this small complex, LAP becomes covalently linked to one of the four latent TGF-β binding proteins (LTBPs), forming the large latent complex. The LTBP has an ECM-binding region that allows targeting into the mesangial ECM (Figure 9) (78).

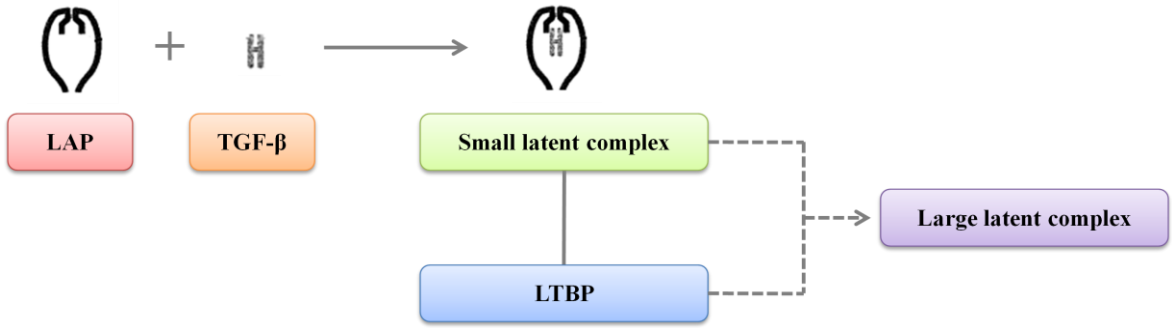

Figure 9. Formation of the large latent complex (adapted from (78)). LAP: latency-associated peptide; LTBP: latent TGF-β binding protein; TGF-β: transforming growth factor β.

This large latent complex is then subjected to proteolysis by multiple proteinases of the serine protease family, being released from LTBP and consequently from the mesangial ECM, resulting in a soluble complex localized at the podocyte surface (77,78). This complex is completely activated by the disruption of the non-covalent interaction between TGF-β and LAP, enabling this growth factor to bind its signaling receptors (78).

The renin-angiotensin-aldosterone system (RAAS) plays an important role in diabetic nephropathy. Its activity is augmented in individuals where this complication is present. In response to hyperglycemia, angiotensin II, the active octapeptide derived from

angiotensinogen in the RAAS, is increased, being this response mediated, in part, by ROS (44,79). The overproduction of ROS can be part of a positive feedback loop responsible for perpetuating the RAAS activity in diabetic nephropathy. Angiotensin II contributes to ROS production by activating NADPH oxidase in the vascular smooth muscle cells and mesangial cells. At the same time, ROS activates the NF-kB, which mediates further angiotensinogen activation and consequently increases angiotensin II (44,77). This latter, along with hyperglycemia, is responsible for the ECM accumulation, mainly by stimulating the synthesis of ECM proteins through an enhanced TGF-β1 production in mesangial cells (44). TGF-β1 gene expression is stimulated through two AP-1 binding sites in the promoter region of the growth factor gene. These binding sites, designated box A and box B, regulate the promoter activity and an increased binding activity of the AP-1 transcription factor complex components, transcription factors JunD and c-FOS, predominantly to the box B, results in increased TGF-β1 promoter region activity and consequently enhanced TGF-β1 expression (44,80,81).

The overexpression of TGF-β1 leads to an imbalance in the ECM metabolism, since its action mechanisms include both increased ECM synthesis by stimulating the production of ECM proteins and inhibition of ECM degradation by inhibiting the synthesis of MMPs, stimulating the production of the MMPs inhibitors, TIMPs, and also by inhibiting the synthesis of collagenases, the enzymes responsible for collagen degradation (44,60,61,72,79). These matrix-related actions of TGF-β1 are mediated by the Smad proteins, a family of intracellular signal transducers that act downstream of receptors for the TGF-β family members (44,82). Members of the TGF-β family, which include TGF-βs, activins and bone morphogenetic proteins (BMPs), bind to type II and type I serine/threonine kinase receptors (83,84). The type II receptor kinases are active and, upon ligand binding, the type II receptors activate type I receptor kinases through phosphorylation of the juxtamembrane domain. These type I receptor kinases then activate intracellular substrates like the Smad proteins (84). Smads can be divided into three subclasses, the receptor-regulated Smads (R-Smads), which are anchored to the cell membrane through membrane-bound proteins, including Smad-anchor for receptor activation (SARA), the common-partner Smads (Co-Smads), which assemble into heteromeric complexes with the R-Smads, and the inhibitory Smads (I-Smads), which function as antagonists of R-Smads and Co-Smads. The R-Smads directly interact with and

become phosphorylated by type I receptors, then forming complexes with Co-Smads and migrating into the nucleus, where they regulate the transcription of target genes by interacting with various DNA binding proteins or by recruitment of transcriptional co-activators or co-repressors (Figure 10). In mammals, TGF-βs/activin-specific R-Smads, Smad2 and Smad3, are phosphorylated by type I receptor kinase and form a heteromeric complex with Smad4, the only Co-Smad in mammals. The complex is translocated into the nucleus, where it regulates the transcription of the target genes (83,84,85).

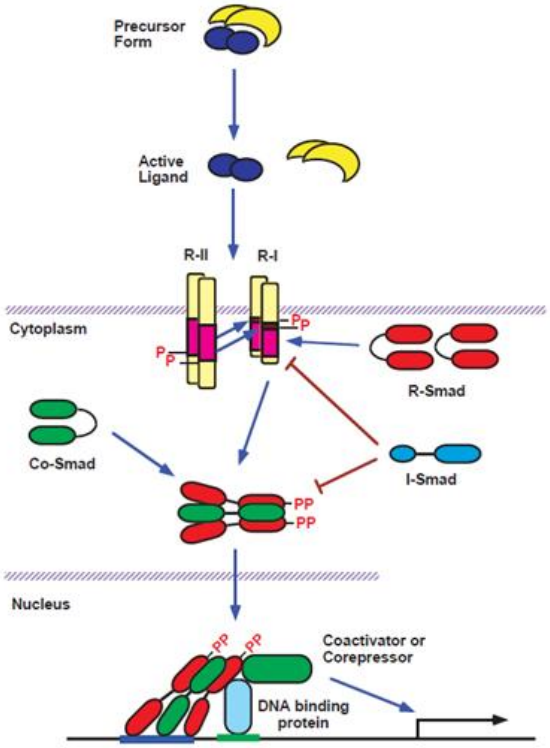

Figure 10. Signaling by TGF-β through serine/threonine kinase receptors and Smad proteins (adapted from (84)). Co-Smad: common-partner Smads; I-Smad: inhibitory Smads; R-I: type I receptor; R-II: type II receptor; R-Smad: receptor-regulated Smads.

The R-Smads also include Smad1, Smad5 and Smad8, that are BMP-specific Smads and the Smad6 and Smad7 are I-Smads (84,85). In a similar way to the R-Smads, the I-Smads interact with type I receptors kinases and compete with the Co-Smads for complex formation with the R-Smads. Smad7 is a potent inhibitor of TGF-β signaling, being its expression directly induced by effects of the Smad3 and Smad4 on its promoter, and the Smad6 preferentially inhibits BMP signaling (84). Although the main signaling pathway of

TGF-β is the Smad signaling pathway, activation of the p38 mitogen-activated protein kinases (MAPK) signaling pathway may also occur (44).

Therefore, and due to the action mechanisms associated with TGF-β1 previously mentioned, the overexpression of this growth factor results in ECM accumulation in the renal tubulointerstitium, the mesangium and the basement membrane of the glomerulus due to excessive deposition of either proteins that are normally present in these structures, or proteins that are not present in the normal tissue, or both (Table 4) (44).

Table 4. Proteins that are normally present in the mesangium and glomerular basement membrane and changes in the accumulation of those proteins in diabetic nephropathy (adapted from (44)).

	Protein	Normal/Diabetic Nephropathy
Mesangium	Collagen I	Only detected in advanced glomerulosclerosis
	Collagen III	Only detected in advanced glomerulosclerosis
	Collagen IV	Chains α1(IV) and α2(IV) expressed in normal mesangium/ Increased in diabetic nephropathy
	Collagen V	Minor component in normal mesangium/ Increased in diabetic nephropathy
	Collagen VI	Present in normal mesangium/ In diabetic nephropathy has the same distribution as α1(IV) in normal mesangium
	Fibronectin	Present in normal mesangium/ Increased in diabetic nephropathy
	Laminin	Minor component in normal mesangium/ Increased in diabetic nephropathy
	SLR Proteoglycans*	Present in normal mesangium/ Only detected in advanced glomerulosclerosis
Glomerular basement membrane	Collagen IV	Chains α3(IV) and α4(IV) present in normal glomerular basement membrane/ Increased in diabetic nephropathy; Chains α1(IV) and α2(IV) are minor components in normal glomerular basement membrane/ Decreased in diabetic nephropathy
	Entactin (Nidogen)	Present in normal glomerular basement membrane/ Increased in diabetic nephropathy
	Laminin	Present in normal glomerular basement membrane/ May be increased in early diabetic nephropathy followed by a decrease
	Heparan Sulfate Proteoglycan	Present in normal glomerular basement membrane/ Decreased in diabetic nephropathy

SLR Proteoglycans: small leucine-rich proteoglycans.
*Includes decorin, biglycan, lumican and fibromodulin

Protein deposition leads to mesangial expansion and glomerular basement membrane thickening, eventually leading to glomerulosclerosis and tubulointerstitial fibrosis (44). The comparison between a normal glomerulus and a diabetic one is presented in Figure 11.

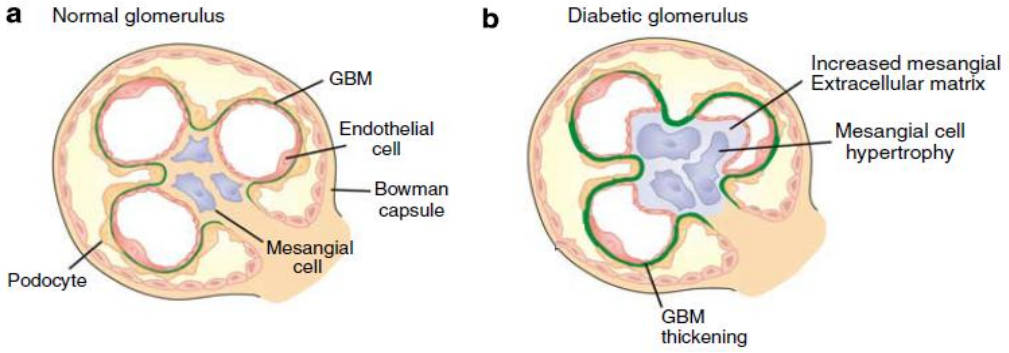

Figure 11. Diabetic nephropathy associated glomerular changes (adapted from (71)). GBM: glomerular basement membrane.

The TGF-β1 active form, localized in the podocytes, has a very short half-life in plasma, making it unlikely for TGF-β1 to transverse the glomerular basement membrane to promote mesangial expansion (77). Instead, TGF-β1 induces the production of CTGF, via Smad binding elements and a unique TGF-β response element in the CTGF promoter, to mediate TGF-β1 stimulated ECM accumulation in the mesangial cells of the mesangium (44,77). The mesangial expansion, due to abnormal ECM metabolism in mesangial cells, reduces the capillary surface area available for glomerular filtration, contributing to the progressive loss of renal function. This structural change in the kidney is frequently associated with the development of diffuse glomerulosclerosis, characterized by a distributed mesangial expansion, or nodular glomerulosclerosis, which happens when there are areas of marked mesangial expansion, forming roundish and fibrillar zones with palisade arrangement of the nuclei, the Kimmelstiel-Wilson nodes (58,86,87). Furthermore, the mesangial expansion also correlates with tubulointerstitial fibrosis, a disease much less well studied than glomerulosclerosis. Even though the mechanisms of tubulointerstitial fibrosis are not clear, it is known that progressive loss of renal function associates with the worsening of the fibrosis (44). The thickening of the glomerular basement membrane, a consequence of the excessive production of the membrane matrix components by the podocytes and endothelial cells, results in modifications of the charge-selective properties of this structure, consequently contributing to proteinuria (49,87).

4.1.2. Dysfunction and/or loss of podocytes

Some of the changes that occur in the kidney of individuals with diabetic nephropathy lead to proteinuria, being this event associated mainly with modifications in the glomerular filtration barrier (86). This barrier is highly permeable to water and small molecules but maintains very low permeability to macromolecules, being these characteristics dependent on its unique three-layer structure (88). The glomerular filtration barrier is, therefore, composed of the fenestrated glomerular endothelium, the glomerular basement membrane, and the glomerular visceral epithelial cells, widely known as podocytes (Figure 12) (45,86,89).

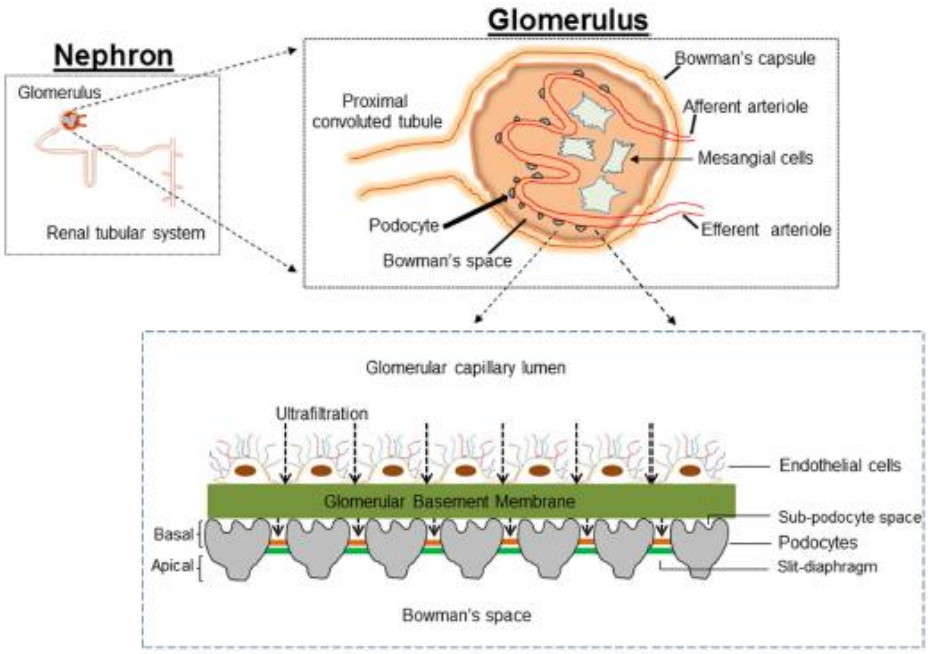

Figure 12. Architecture of the glomerular filtration barrier in the kidney glomerulus (taken from (45)).

The first layer encountered in the glomerular filtration barrier starting from the glomerular capillary lumen, is the fenestrated glomerular endothelium (Figure 13). This endothelium consists of endothelial cells characterized by fenestrations, round or ovoid transcellular holes through the cytoplasm with 60-80 nanometers (nm) of diameter (88,90). Until recently, it was thought that the fenestrated endothelium contributed little to the selective permeability of the glomerular filtration barrier (88). However, it is now known that the endothelium is covered with glycocalyx, a negatively charged hydrated mesh of

cell-surface anchored proteoglycans, adsorbed proteins and glycosaminoglycans in dynamic equilibrium with the circulating plasma. The glycocalyx forms a significant permeability barrier, making it possible for the fenestrated endothelium to present some restriction to the passage of plasma proteins (86,88,90).

Separating the fenestrated endothelium from the podocytes is the second layer of the glomerular filtration barrier, the glomerular basement membrane (Figure 13). This structure provides support for the glomerular capillaries and represents the ECM component of the glomerular filtration barrier. As part of this barrier, the glomerular basement membrane is considered to be a "crude prefilter" that functions as a size and charge-selective filtration barrier, since macromolecules as well as negatively charged molecules are impaired from crossing it. While the size-selectivity is related to the porosity of the membrane conferred by the laminin and collagen IV, the major components of the glomerular basement membrane, in concert with the remaining membrane components, the charge-selectivity of this structure can be attributed to agrin, the major heparin sulfate proteoglycan present in the membrane. Agrin has sulfated glycosaminoglycan side chains that confer to it a highly negative charge, being, therefore, responsible for the anionic sites present within the basement membrane (49,86).

Lastly, the podocytes are the final layer of the glomerular filtration barrier, functioning as the ultimate size-selectivity barrier and only permitting permeability to molecules smaller than albumin (Figure 13) (87). They are highly specialized epithelial cells that cover the urinary side of the glomerular basement membrane (45). The podocytes are comprised of a cell body and cytoplasmic extensions called major processes, which are divided into actin-rich foot processes that interdigitate and cover the capillary walls of the glomerulus (45,91). The podocyte foot processes are a contractile system similar to that seen in pericytes, and, while podocyte major processes are constituted by microtubules and intermediate filaments, the foot processes present a complex mesh of actin filament bundles with a dense network of F-actin and myosin. These actin filament bundles form arches between the foot processes of the same podocyte, constituting the actin cytoskeleton. The cytoskeleton is connected to the underlying glomerular basement membrane via the α3β1 integrin and α- and β-dystroglycans, consequently anchoring the foot processes to this membrane and maintaining the podocyte integrity, a requirement for a fully functional glomerular filtration barrier (45,87,92,93). At the interdigitation site, the

foot processes from neighboring podocytes are connected to each other by an intercellular adherent junction, the slit diaphragm, which is located just above the glomerular basement membrane (48,91). The slit diaphragm consists of several proteins that, besides forming the protein complex that contributes to its structure, also connect the diaphragm to the actin cytoskeleton, which ultimately determines the structural maintenance of the slit diaphragm, and enable it to do its function as the structure responsible for the podocytes size-selective permeability (Figure 13) (45,89,93).

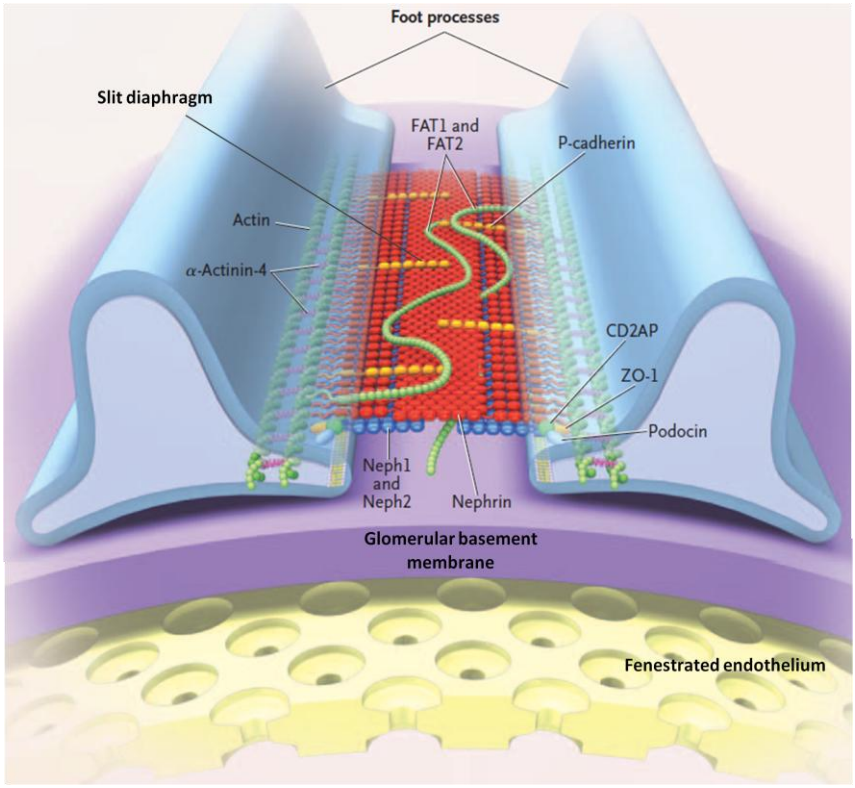

Figure 13. Protein components of the podocyte slit diaphragm (adapted from (89)). CD2AP: CD2-associated protein; FAT1: FAT atypical cadherin 1; FAT2: FAT atypical cadherin 2; Neph1: nephrin-like protein 1; Neph2: nephrin-like protein 2; P-cadherin: placental cadherin; ZO-1: zonula occludens-1.

Nephrin is a large transmembrane molecule, being its N-linked glycosylation critical for the plasma membrane localization of this molecule. Its predicted structure and biochemical properties suggest that nephrin may form dimers across the filtration slit, meaning that nephrin molecules from adjacent foot processes are thought to interact with one another in the middle of the slit diaphragm to form a central density that functions as a filtering

structure, since it contains 30-40 nm wide pores, the filtration slits (77,89,93). The proteins nephrin-like protein 1 (Neph1) and nephrin-like protein 2 (Neph2) are structurally related to nephrin, forming heterodimers with it that can transduce signals from the podocyte slit diaphragm that lead to actin filament polymerization, and, therefore, contributing to the stabilization of the diaphragm (89,92). The podocin, a membrane protein from the stomatin family, is located solely in the slit diaphragm region and, like other stomatin family members, recruits its complex partners to cholesterol-enriched membrane domains in the slit diaphragm (lipid rafts - specialized membrane domains enriched in cholesterol, glycosphingolipids and GPI-anchored proteins critical for the dynamic functional organization of the slit diaphragm), thereby creating a cluster that can act as a unique signaling platform. Podocin associates via its C-terminus with nephrin and neph1, forming the nephrin-neph1-podocin receptor complex. This complex interacts with proteins associated with the actin cytoskeleton, like CD2-associated protein (CD2AP), signaling to regulate cytoskeleton dynamics and morphology, as well as actin remodeling (89,92,93). The placental cadherin (P-cadherin), as well as FAT atypical cadherin 1 (FAT1) and FAT atypical cadherin 2 (FAT2) proteins, are widely expressed cadherin superfamily proteins that provide structural support to the slit diaphragm (93). Other proteins such as CD2AP protein is localized at the intracellular insertion site of the slit diaphragm and interacts with nephrin, functioning as an adaptor that connects this protein to the actin cytoskeleton. While CD2AP does not appear to be necessary for podocyte development, it is essential for maintaining podocyte structure, since an efficient association between the slit diaphragm and the actin cytoskeleton is required for the stabilization of the diaphragm (89,92,93). The protein zonula occludens-1 (ZO-1), similarly to CD2AP, is also localized at the intracellular insertion site of the diaphragm and interacts with actin filaments, but its precise role in the protein complex is still unknown (89,93). The last protein involved with the actin cytoskeleton is α-actinin-4. This protein is an actin filament cross-linking protein that cross-links F-actin filaments in the foot processes, contributing to the normal assembly or disassembly of actin filaments in podocytes (45,87,89). Besides this, α-actinin-4 also interacts with components of the integrin complex at the glomerular basement membrane, more specifically the α3β1 integrin, making the connection between the foot processes and the glomerular basement membrane (93).

The glomerular filtration barrier functions as a whole, meaning that disruption of any of the layers affects the overall permeability of the barrier, which can result in proteinuria (88). The role of each component of the glomerular filtration barrier in the development of proteinuria in diabetic nephropathy has been subjected to much debate. The dysfunction of the fenestrated glomerular endothelium as well as the thickening of the glomerular basement membrane, which, as mentioned before, leads to the loss of its charge-selective properties, have been proposed to explain partly the pathogenesis of proteinuria (45). The glomerular basement membrane charge-selectivity is, as previously mentioned, a property attributed to agrin, a heparin sulfate proteoglycan, and its loss can be explained by the presence of hyperglycemia as well as the increase in angiotensin II. The presence of hyperglycemia suppresses the production of the agrin core protein while angiotensin II, in addition to decreasing the synthesis of the agrin core protein, also diminishes the sulfation of its side chains, causing it to lose its negative charge (86). However, this loss of charge-selective only occurs late in the course of diabetic nephropathy, sometimes long after the development of microalbuminuria, which suggests that other glomerular filtration barrier components may have a more pivotal role in the pathogenesis of proteinuria in diabetic nephropathy (45,86). These observations along with the data collected over the past decade has highlighted the crucial role of podocytes in the glomerular filtration barrier, more specifically, the crucial role of the podocytes slit diaphragm in the filtering process, being the integrity of the slit diaphragm what ultimately determinates the permeability properties of the glomerular filtration barrier (45,86,94).

Podocytes are old cell types, being expected that their features have been extensively shaped by natural selection. Their complex structure and their unique position as “floating” cells fixed to the glomerular basement membrane only by their foot processes are somehow indispensable for their function (95). In diabetic nephropathy early podocyte damage occurs, resulting in podocyte dysfunction, and causing the podocytes to undergo tremendous changes in their shape without losing its viability (86,95). The most general change is the loss of the interdigitating foot process pattern, usually called foot process effacement. This effacement is a consequence of the retraction, widening or shortening of the foot processes of each podocyte, being associated with the leakage of macromolecules, such as albumin, through the glomerular filtration barrier (87,95). The foot process effacement can be caused by interference with the slit diaphragm protein complex,

interference with the glomerular basement membrane or with its interaction with the podocyte or it can also be caused by abnormalities in the actin cytoskeleton, being this phenomenon separated into two stages (87,93,95). Within the first stage, foot processes undergo tremendous changes in shape, losing their regular interdigitating pattern and retracting into short irregularly shaped cells, causing the slit diaphragms to be lost or displaced from their usual position at the base of the processes. The second phase includes retraction of the foot processes into the primary podocyte foot processes, leading to broad flattened disc-like projections covering the glomerular basement membrane (95). In addition to this, decreased expression of some proteins associated with the slit diaphragm, such as nephrin, P-cadherin, and ZO-1, impairs the integrity of this structure, correlating with the development of foot process effacement and changes in podocyte permeability. Furthermore, besides a diminished expression, nephrin localization in the podocyte foot processes is also altered in individuals with diabetic nephropathy, which also contributes to the impairment of the slit diaphragm integrity and the development of foot process effacement (45,87). Therefore, foot process effacement is not a simple pathological derangement of normal podocyte architecture, but it seems to represent a regulated change in the podocyte that can be considered as a reestablishment of a more typical epithelial cell structure. The detailed mechanisms associated with this phenomenon are still poorly understood (95).

In diabetic nephropathy, the number of podocytes is reduced (podocytopenia), being this reduction correlated with disease progression (86). Usually, podocyte reduced numbers reflect the imbalance between podocyte proliferation and podocyte loss (87). The podocytes exit the cell-cycle and remain terminally differentiated, lacking a proliferative mechanism in response to injury (45,87,93). Meanwhile, the exact etiology of podocyte loss remains unclear, but two different mechanisms can be suggested for it: apoptosis and detachment (86). Apoptosis is regarded as a mechanism by which podocytes are lost. Apoptosis of podocytes occurs under very specific conditions that can develop during diabetic nephropathy progression, such as the overexpression of TGF-β1 (95). This growth factor induces apoptosis through Smad pathways, specifically Smad7, and via p38 MAPK (45,77,87,93). Even though the full mechanisms are still not clear, it is known that Smad7 is responsible for podocyte apoptosis by inhibiting the nuclear translocation and transcriptional activity of a survivor cell factor, the NF-kB (45,86). The podocytes unique

situation by being “floating” cells only connected to the glomerular basement membrane by their foot processes and being exposed to the flow of filtrate, makes them highly susceptible to detachment. In diseases where there are changes in the composition of the glomerular basement membrane, such as diabetic nephropathy, a weaker adhesion of podocytes to this membrane is expected to occur, due to both the changes in the glomerular basement membrane itself as well as changes in the proteins that connect these two structures. The main mechanism associated with the detachment process is interference with components of the integrin complex, more specifically with α3β1 integrin (95). The overexpression of the growth factor TGF-β1, in addition to inducing podocyte apoptosis, also contributes to a reduction in α3β1 integrin expression, the integrin responsible for establishing the connection between the foot processes and the glomerular basement membrane (77). Therefore, an impaired podocyte adhesion to the underlying glomerular basement membrane may result from a downregulation of α3β1 integrin (77,86). Besides α3β1 integrin, it is also thought that changes in dystroglycan expression contribute for podocyte detachment. However, those changes under diabetic conditions have not been fully explored yet (87). Furthermore, podocyte detachment may occur in conjunction with apoptosis or podocytes can detach as viable cells (86,95). As a consequence of the detachment, detached podocytes are present in the urinary sediment of individuals with kidney diseases, being the majority of those detached podocytes still viable, with a well-formed nucleus with a normal chromatin pattern and a cytoplasm with well-preserved organelles. For that reason, it is thought that the majority of podocytes that go through detachment are still viable, and, during that process undergo changes in their shape, being foot process effacement frequently present. Therefore, foot process effacement appears to precede podocyte detachment (95).

With the increased loss of podocytes and their limited proliferative capacity, a critical proportion of the podocyte population of the glomerulus is lost, being the remaining podocytes unable to compensate for the lost ones, resulting in an impaired glomerular filtration function (45,77). These factors, along with the mesangial expansion, result in the development of glomerulosclerosis (45,77,93). This development occurs firstly because of podocyte loss, since the inability to replace the lost podocytes result in a localized “bare” or denuded glomerular basement membrane at that site. The lack of tensile support normally provided by podocytes is lost in the denuded area of the basement membrane,

which leads to the outward bulging of the capillary loop due to capillary hydrostatic pressures. This expanding capillary loop causes the glomerular basement membrane to come into contact with the Bowman's capsule, resulting in synechia formation (86,93). Finally, subendothelial hyalinosis develops in the affected capillary, and progressive scarring ensues (87,93).

In addition to the effects already mentioned, the overexpression of TGF-β1 also stimulates the expression of vascular endothelial growth factor (VEGF) in podocytes (58,86). This factor acts through VEGF receptors (VEGFR)-1 to 3 and comprises isoforms from VEGF-A to VEGF-E, being its strong permeability properties associated with increasing permeability of the glomerular filtration barrier to plasma proteins (86,90). Podocytes not only produce this factor but are acted upon by it through the VEGFR-1. The VEGFA stimulates the podocyte production of the chain α3 (IV) of the collagen IV, a major component of the glomerular basement membrane, having, therefore, a role in the regulation of this membrane composition by podocytes and contributing to the thickening of the glomerular basement membrane (86). Furthermore, VEGF-A also moves across the glomerular basement membrane and connects to VEGFR-2 on the fenestrated glomerular endothelium, being critical for the maintenance of the endothelium, since it is a key inducer of its fenestrations (58,88,90). Ultimately, this factor induces endothelial nitric oxide production, thereby promoting the vasodilation and glomerular hyperfiltration that are typical of early diabetic nephropathy (58,86).

4.2. Disease management and treatment

Although there is no definitive cure, adequate treatment like a strict glycemic control, early treatment of systemic hypertension, dietary protein restriction, and lipid-lowering therapy can delay the progression of nephropathy (96). Patients that present this complication show a progressive decline in glomerular function and, the treatment of systemic hypertension in these individuals slows down the rate of renal function loss (97). The most effective way for delaying its progression is the blockade of the RAAS using angiotensin-converting enzyme (ACE) inhibitors and angiotensin receptor blockers (ARBs). This treatment, in addition to interrupting the RAAS at different levels, also lowers the blood pressure and provides renoprotective effects, possibly because of the

decrease in glomerular capillary hydraulic pressure due to the removal of the vasoconstrictor effect of angiotensin II in the efferent arteriole (58,97,98,99).

5. Genetic models of complex diseases

Many human complex diseases and related traits are believed to be influenced by several genetic and environmental factors, but until recently, the identification of genetic variants that contribute to these complex diseases has been slow (100). In recent years, GWAS have been extensively used to uncover the genetic architecture of common complex diseases. This method is based on the "common disease, common variant" hypothesis, where allelic variants present in more than 5% of the population are responsible for the development of the disease. However, part of the genetic contribution to complex traits remains unexplained since most common variants individually or in combination, only confer small increments in disease risk and only explain a small proportion of disease heritability (the portion of phenotypic variance in a population explained by additive genetic factors), creating the so-called missing heritability problem (100,101).

Several explanations have been suggested to explain the missing heritability, including a larger number of small effect variants, the rare variants, yet to be found. The rare variants possess a cumulative strong biological effect and are not detected by the available methods, since they only focus on variants present in 5% or more of the population, the structural variants are poorly captured by the existing arrays and there is a current low power to detect gene-by-gene and gene-by-environment interactions (100,102). Much of the speculation about the current missing heritability focuses on the possible contribution of rare variants since these variants could have substantial effect sizes without demonstrating clear Mendelian segregation and could explain additional disease risk or trait variability. Besides this, also the evolutionary theory predicts that deleterious alleles are likely to be rare as a result of purifying selection, and indeed, variants that cause loss of function and, therefore, prevent the generation of functional proteins, are especially rare (100,101).

Given that little has been explained about most complex diseases despite the identification of hundreds of associated genetic variants, the search for the missing heritability can provide a potentially important path towards further discoveries (100).

6. Methodologies to obtain genetic information of complex diseases

As previously stated, it is now known that complex diseases are, in part, determined by genetic factors, which resulted in increased interest in identifying the genetic basis of those diseases (19). For that reason, during the past few years several methods were used to study and identify genes that could be associated with complex diseases (9).

6.1. Genome-Wide Association Studies

The genome-wide association studies, known as GWAS, have discovered multiple genetic variants that can be associated with an ever increasing number of diseases (103). This method is based on linkage disequilibrium (adjacent polymorphisms are correlated with each other due to their co-segregation from one generation to the next) and advanced genotyping techniques (microarrays through which many tag SNPs can be typed in a single array) (40). This technique is used to find connections between genes and diseases across a population since it uses a database with over a 1,000,000 known common variants. Therefore, this methodology, although it requires large sample sizes to detect effects of common variants (minor allele frequency (MAF) > 5%) with genome-wide significance, can simultaneously analyze several variations, helping to identify dozens of new associations between genes and diseases (9,40,104).

GWAS has allowed to deepened our understanding of disease etiology in multiple directions because, first, it has led to the identification of a vast number of disease-associated genomic loci/genetic variants and secondly, the associated loci/variants discovered from GWAS could serve as risk predictors for some diseases, provided large enough GWAS discovery sample size. An example of this can be seen in the study of Crohn's disease, an inflammatory bowel disease with a prevalence of 0.32% in Europe and North America. GWAS identified 163 loci at genome-wide significance level, which collectively explain 13.6% of this disease phenotypic variance. Moreover, valuable insight can be learned about the genetic etiology of many diseases due to the analysis of the

variance contribution of SNPs from certain genomic regions and pathways. However, GWAS still presents problems such as the difficulty in interpreting its results partly because of our limited understanding of the genomic function, especially the non-coding regions in which a considerable number of genetic variants been identified, and because of the correlation structure among neighboring variants, often referred to as linkage disequilibrium (102).

Even though the systematic identification of rare variants associated with complex diseases is not yet done, several of those variants have already been identified as conferring risk for disease development (for example in autism, mental retardation, epilepsy and schizophrenia). It is clear that the common variants found by GWAS have little success in implicating specific genes in specific diseases, which ultimately reduces their importance in several applications such as drug development (104). Due to the limitations of GWAS and to investigate the rare genetic variants that are not captured by this method, whole-genome and whole-exome sequencing began being used as a substitute for the screening of genes (103,104).

6.2. Whole-Exome Sequencing

Whole-exome sequencing (WES) is the method by which only the coding regions of the genome (1-2%), the exome, are sequenced, enabling the detection of the majority of important pathogenic variants (101,105,106). This method targets the consensus coding sequence (of the collaborative consensus coding sequence (CCDS) Project) (107), which is approximately 30,000,000 bases, only differing the targeted regions depending on the service provider (101). This technique has been, therefore, proposed to provide a new strategy to study the missing heritability (108). In recent years, the emergence of exome sequencing has led to a change in the approach to identify new causal mutations and genes (103). The current ability for interpretation of the functional consequences of sequence variation outside the coding regions (introns) is limited, and exome sequencing became an efficient strategy for identification of rare functional variants (101,109). Additionally, WES has recently begun being explored as a diagnostic tool for genetic diseases (103).

The utility of exome sequencing as a diagnostic tool is becoming evident for disorders characterized by genetic heterogeneity. Diseases characterized by locus heterogeneity can

result from a causal mutation on one of several candidate genes, which results in the need for each one of those genes to be screened in its entirety by polymerase chain reaction (PCR) and Sanger sequencing, a laborious and expensive process. On the same way, also phenotypic heterogeneity interferes with the accuracy of the assessment of clinical manifestations, resulting in a potentially incorrect or ambiguous diagnosis. As a result, only the more probable causal genes are prioritized for screening. An example of this is in disorders caused by mutations in several candidate genes such as retinitis pigmentosa and Charcot-Marie-Tooth disease, for which the diagnostic strategy involves gene by gene screening. (103). In response to this limitation, the exome sequencing began to be used as a diagnostic tool and its advantages are evident comparing to the traditional approach (103,110). For cases that are not explained by mutations in known loci, the distinction between the diagnostic tool and the discovery one is a little blurred. A definitive genetic diagnosis is not made based only on the new found mutation; screening in additional cases is required, and the discovery of the new mutation in those cases from different families is the only confirmation of causality (103).

In exome sequencing, the workflow is divided into three phases: sample preparation and sequencing, primary data processing and secondary data processing (Figure 14).

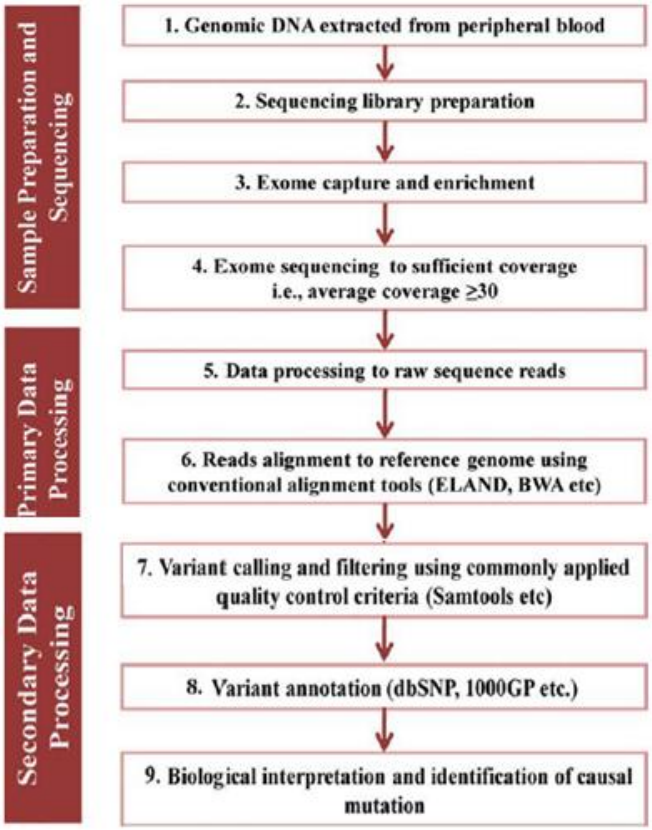

Figure 14. Exome sequencing workflow since genomic DNA extraction to biological interpretation and identification of the causal mutation (adapted from (103)). 1000GP: 1000 genome project; BWA: burrows-wheeler aligner; dbSNP: single nucleotide polymorphism database; DNA: deoxyribonucleic acid.

The library preparation uses genomic DNA and involves several steps like DNA fragmentation and adapter ligation. Exon capture and enrichment are usually done using commercial kits before proceeding to exome sequencing with one of the high-throughput next-generation sequencing (NGS) platforms (steps performed according to the manufacturer's recommended protocols). According to the platform used, there is a minimum value of coverage considered sufficient for the correct calling of variants. In recent years, several technical approaches have been developed to enrich selectively for the protein-coding regions of the genome using modified multiplex PCR, capture by circularization or capture by hybridization in solution or on an oligonucleotide array. After sequencing, the data obtained is processed to raw sequence reads, which are then aligned to a reference genome. Finally, variant calling and filtering is done using quality control criteria followed by variant annotation to obtain information such as genomic position and

functional effect, and after, those variants are analyzed to find the causal mutation (103,111).

Despite all the advantages, in practice, various improvements in WES are yet to be introduced. Most importantly, WES almost entirely misses structural variation, only small (1-25 nucleotides) insertions and deletions can reliably be detected and, if causal variations are in exons not targeted, such as novel exons in known and novel genes, they will not be identified. Another major challenge is to analyze and manage the large amounts of sequencing data generated since exome sequencing usually generates much more than 10,000 genetic variants that have to be filtered to identify only the causal mutations. Also, since the success of WES depends on the capture of all known exons (approximately 200,000) followed by an adequate depth of sequencing, the methods for exome capture, even though suitable, can be improved (103,104,106).

Because for many applications it is not necessary the sequencing of entire genomes and the most disease-causing variants are in the coding regions (approximately 85% in Mendelian diseases), WES is a cost-effective alternative to GWAS or even whole-genome sequencing (WGS), since this latter still presents challenges concerning technical, analytical and economic issues (103,104,105). At the present, the introduction of NGS has allowed fast and relatively inexpensive sequencing of thousands of genes at a time through WES (19). Therefore, with the affordability with WES and with the advances in NGS technologies, researchers are gaining the genomics tools to discover associations between coding variants and complex diseases and directly test the hypothesis that rare variants may be associated with increased risk of developing complex diseases (19).

Regarding ethics, the unmasking of genetic details has always been intertwined with some complex ethical implications (111). With the increasing capability of sequencing genomes, new ethical issues were raised (105). For certain diseases where the causal mutations are already known, the information on whether the individual under study has or not the mutation is available but, when that is not the case, should the researchers have the obligation to screen for other variants besides the ones related to the disorder being investigated? Several positions are in discussion, being one of them that researchers do not have the obligation to screen for other variants besides the ones related to the disorder being investigated, because, among other things, that obligation would damage the researcher ability to focus on the particular disease being studied. Meanwhile, the

international research and clinical community are still discussing the ethical implications of the information availability (112).

6.2.1. Next-Generation Sequencing

Different methods to sequence DNA were developed in the 1970s. Sanger and colleagues developed the use of chain-terminating dideoxynucleotide analogs that caused base-specific termination of primed DNA synthesis (113). In an alternative approach, Maxam and Gilbert suggested a method in which labeled DNA fragments were subjected to chemical cleavage at specific bases and the reaction products were separated by gel electrophoresis (114). Since the method developed by Sanger and colleagues, commonly referred to as Sanger sequencing, had less handling of toxic chemicals and radioisotopes than the Maxam and Gilbert's technique, it became the obvious choice for DNA sequencing and clinical application involving the analysis of single genes with limited polymorphisms (105,111). Later, Sanger sequencing was improved to provide an increased throughput that led to laboratory automation and process parallelization (105). This progress eventually made it possible to complete the first human genome sequence in 2004 (115) even though it also made clear the need for faster and higher throughput and cheaper technologies (105).

To overcome the limitations of the Sanger method (considered first-generation technology), NGS technologies were developed and commercialized. Those techniques sequence by either creating micro-reactors and/or attaching the DNA molecules to be sequenced to solid surfaces or beads, which results in millions of reactions happening in parallel (105,116). This technology is, therefore, a high-throughput, massively parallel technology (106). Therefore, the appearance of NGS satisfied the growing demand for low-cost sequencing technology producing massive parallel nucleotide sequencing but it also created demands on the bioinformatics tools for data analysis and management (105,117). With this, part of the increase in throughput in NGS is due to advances in data handling (105). The number of reads produced by NGS revolutionized genomics research, however, a drawback to the use of these technologies is the relatively short reads (35-500 base-pairs (bp), depending on the platform) compared with the traditional sequencing

(1,000-1,200 bp), which ultimately made genome assembly more difficult and required new alignment algorithms (105,118,119).

The NGS technologies present major improvements compared with the Sanger sequencing. First, the DNA sequencing libraries are clonally amplified *in vitro* instead of requiring bacterial cloning of DNA fragments; secondly, DNA is sequenced by synthesis rather than chain termination chemistry; thirdly, instead of hundreds, thousands to millions of sequencing reactions are produced in parallel, and finally, the sequencing output is detected directly without needing an electrophoresis. These technological improvements contribute to the growing interest in the use of NGS as a diagnostic tool (105) as well as in other fields (Figure 15) (119).

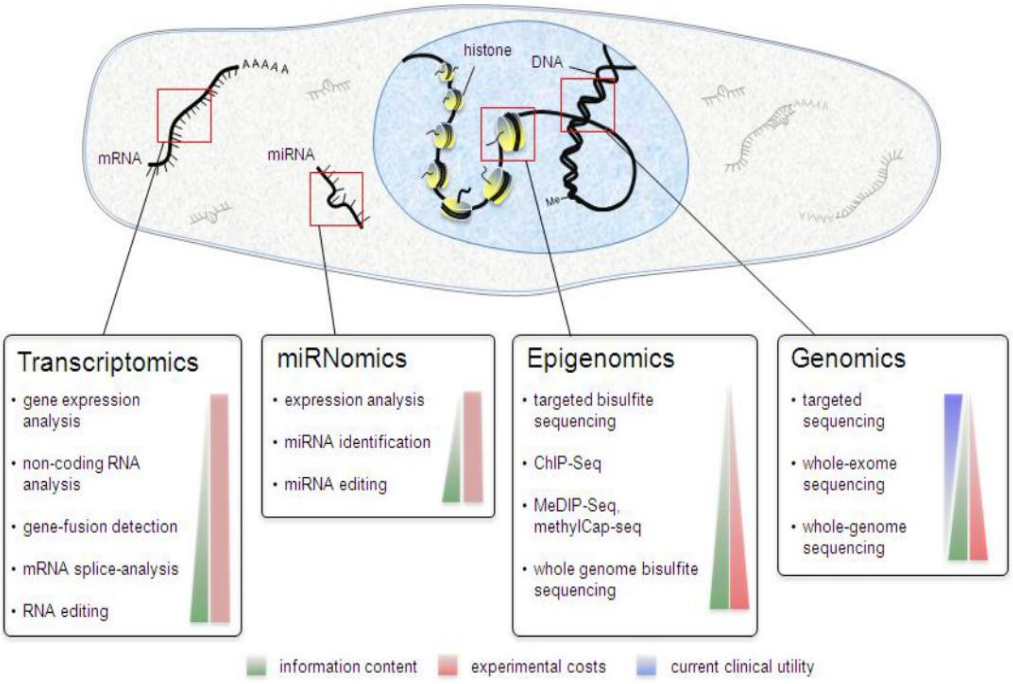

Figure 15. The several NGS applications and the different methods used for which field (taken from (119)). ChIP-Seq: chromatin immunoprecipitation sequencing; DNA: deoxyribonucleic acid; MeDIP-Seq: methylated DNA immunoprecipitation; miRNA: micro ribonucleic acid; mRNA: messenger ribonucleic acid; RNA: ribonucleic acid.

Now, despite the benefits of NGS technologies, a number of technical challenges appeared, such as the storage of information since the technology generates massive volumes of data, and the bioinformatic analysis since the generated DNA sequences have to be aligned to a reference genome, and the short reads may result in difficulties with the

assembling of the generated sequence. Also, both NGS and bioinformatic tools are unable to identify accurately certain genetic variations, such as copy number variations (CNV), inversions, translocations and nucleotide repeat expansions; DNA regions enriched with guanine-cytosine (GC) content that tend to have low coverage. Finally, all NGS platforms have sequencing errors particularly towards the end of the read (106,117,120).

The first NGS technology to be released was the 454 Life Sciences (now Roche) in 2005. This technology uses a pyrosequencing method and currently generates about 1,000,000 reads of 1,000 bp. In 2006, the Illumina Genome Analyzer platform was also released followed by the Sequencing by Oligo Ligation Detection (SOLiD) by Applied Biosystems (now a part of Life Technologies™) in 2007. Even though both of these sequencers generated a greater number of reads than 454 (30,000,000 and 100,000,000 reads, respectively), the length of the reads was only 35 bp long. In 2010, Ion Torrent (now a part of Life Technologies™) commercialized the Personal Genome Machine (PGM). This system was developed by Jonathan Rothberg, the founder of the 454, and it resembles the 454 system. The major difference between the 454 and PGM is that the latter uses semiconductor technology for nucleotide incorporation detection, not relying on the optical detection using fluorescence and camera scanning. As a result, PGM has a higher speed, lower cost and smaller instrument size than 454. Later, from the same company of PGM, Ion Proton emerged (105,121).

Other NGS platforms developed that did not dominate the market were the Qiagen-intelligent bio-systems sequencing-by-synthesis, Polony sequencing and a single molecule detection system (Helicos BioSciences). This latter method, since the template DNA is not amplified before sequencing, is at the interface between NGS and third-generation sequencing technologies (105). Ideally, third-generation sequencing combines all the advantages of NGS with longer read length and decreases in the computational power required to assemble genomes (105,111). The Pacific Biosciences RS (PacBio) platform, a third-generation sequencing technology from Pacific Biosciences, is the first sequencing technology to offer long read lengths (average of 2-3 kilobase (kb)) without GC bias or systematic errors (122). This technology consists of a process in which a DNA polymerase molecule, bound to a DNA template, is attached to the bottom of a well. Each polymerase synthesizes the second strand DNA in the presence of γ-phosphate fluorescently labeled nucleotides and, as each base is incorporated, a distinct fluorescence is detected in real

time (real time sequencing). Besides this, the Pacific Biosciences RS platform offers an amplification-free workflow (123).

6.2.1.1. Ion Proton

Ion Proton is used mainly for sequencing exomes and ribonucleic acid sequencing (RNA-Seq) (121). The basic concept of this sequencer is to perform sequencing-by-synthesis, with electrochemical detection. One of the main products of DNA synthesis is the release of a hydrogen ion (H^+) from the 3' OH incorporation site on the strand being synthesize, and a sensor capable of detecting this product (124).

In Ion Proton, DNA molecules are connected, at both ends of each fragment, to adaptors that contain the necessary elements for immobilization on a solid surface and sequencing. The DNA fragments bound to the adaptors are then captured on the surface of beads. The complex DNA fragment-bead is isolated in an individual water-in-oil emulsion (emulsion polymerase chain reaction (ePCR)) that also contains PCR reagents, and subsequent thermal cycling produces about 1,000,000 copies of each DNA fragment present on the surface of each bead (Figure 16). Only little input of DNA (a few micrograms at most) is needed to produce the library (105,118,125).

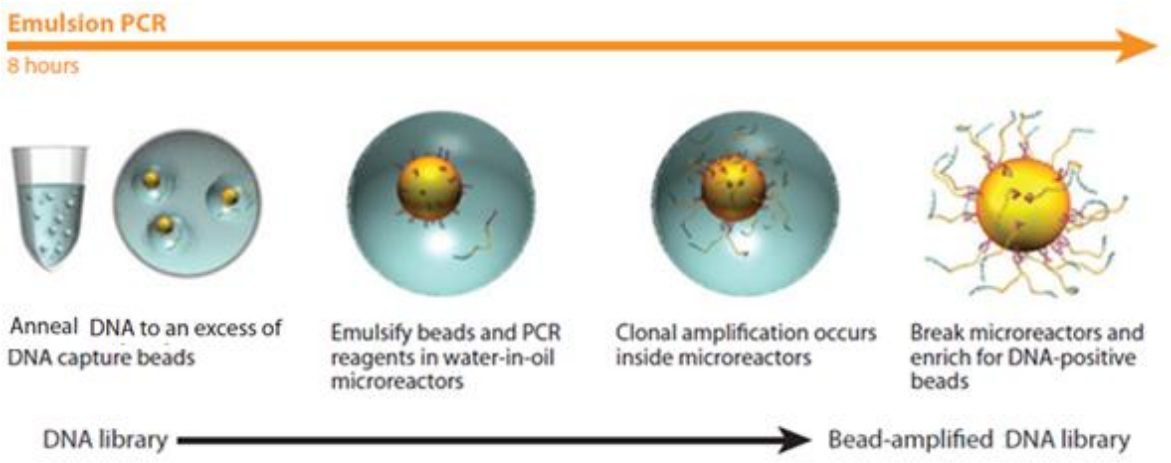

Figure 16. Basic concept of emulsion PCR (adapted from (125)). DNA: deoxyribonucleic acid; PCR: polymerase chain reaction.

The main issues with library preparation are the introduction of quantitative biases and the loss of material. For this purpose, and because PCR is the major source of bias, comparisons between PCR polymerases have been performed, and the enzymes that

introduce less noise have been identified. Regarding the loss of material, DNA fragmentation, end-repair, and adaptor ligation have been combined into a single reaction and gel and column purification steps have been replaced by magnetic beads (105).

After the amplification process, these single molecules are sequenced (125). In the Ion Proton platform, the release of the ion H^+ causes a variation in pH, and that change is detected by a transistor. A practical transistor for sensing pH variation is the pH-sensitive field effect transistor (pHFET), used in solid-state pH meters. This transistor opened the general field of ion sensitive field effect transistor sensors, known as ISFET. The ISFET senses a variety of ionic concentrations being used to serve as a pHFET in Ion Proton. In a simple way, the detection process occurs because the H^+ concentration creates a positive voltage near the gate region of the transistor and, as a result, the current flowing through the transistor changes (Figure 17A). A ISFET sensor is deposited at the bottom of a microwell, and a bead carrying the sample DNA fragment, amplified by ePCR, to be sequenced is placed on top of the ISFET. Each bead normally carries hundreds of thousands of copies of the same DNA fragment to amplify the chemical signal produced by the sequencing reaction and the well acts as a retainer of the bead and helps to localize and retain the H^+ released when the incorporation occurs (Figure 17B). To achieve the massive parallelism, an array of millions of the transistor sensors are formed on a chip (124).

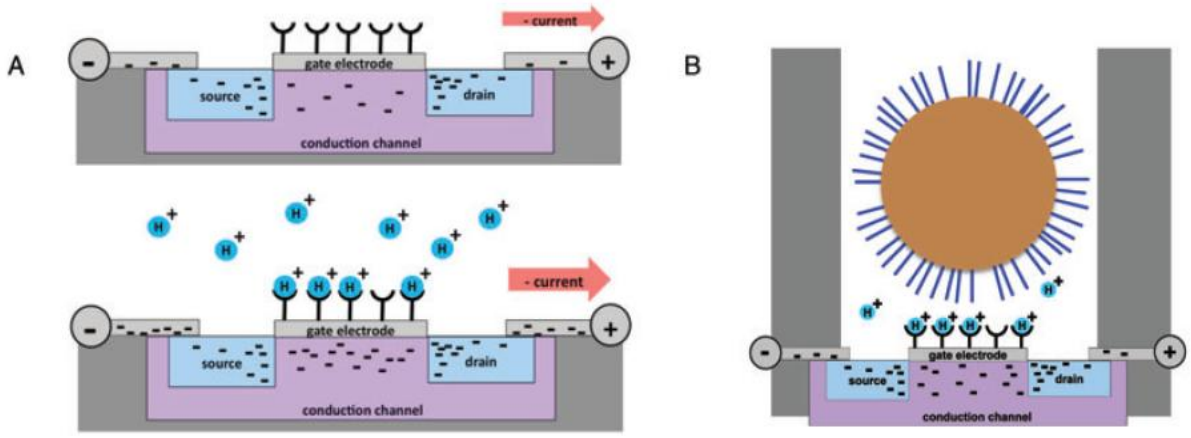

Figure 17. Basic concept of how a transistor works (A) and how that transistor is incorporated into a microwell to detection for the sequencing process (B) (adapted from (124)). H^+: hydrogen ion.

For the sequencing process itself, the DNA fragments that are single stranded and connected to the primers are loaded with polymerase and the sequencing process occurs by sequentially flowing in different nucleotide (adenine (A), cytosine (C), guanine (G) and

thymine (T)) solutions (solution of native deoxyadenosine triphosphate (dATP), deoxycytidine triphosphate (dCTP), deoxyguanosine triphosphate (dGTP) and deoxythymidine triphosphate (dTTP)) and monitoring for an incorporation signal via the sensor. The flowing in of the nucleotide solutions can result in incorporation or no incorporation. If the nucleotide solution does not result in incorporation, the sensor response will be only to whatever change in pH is present in the ambient solution flowing in, which ideally would be none since all solutions should have a matched pH. If, on the other hand, the incorporation occurs (for example if it has been used a solution of dATP and the next base on the template is a T), H^+ ions will be released due to the polymerase activity when incorporating a base and a signal emerges. In this way, the flowing of nucleotide solutions in a repeatedly sequential manner (A, C, G, T, A, C, G, T,...) results in a sensor signal each time a base is incorporated. Between each trial flow, a wash flow is performed to eliminate the previous trial nucleotide solution before moving on to the next one (124).

6.2.2. Bioinformatics Analysis

After sequencing, the initial data analysis pipeline consists of raw data collection and signal processing, followed by base calling, which results in a FASTQ file (Figure 18) (124,126).

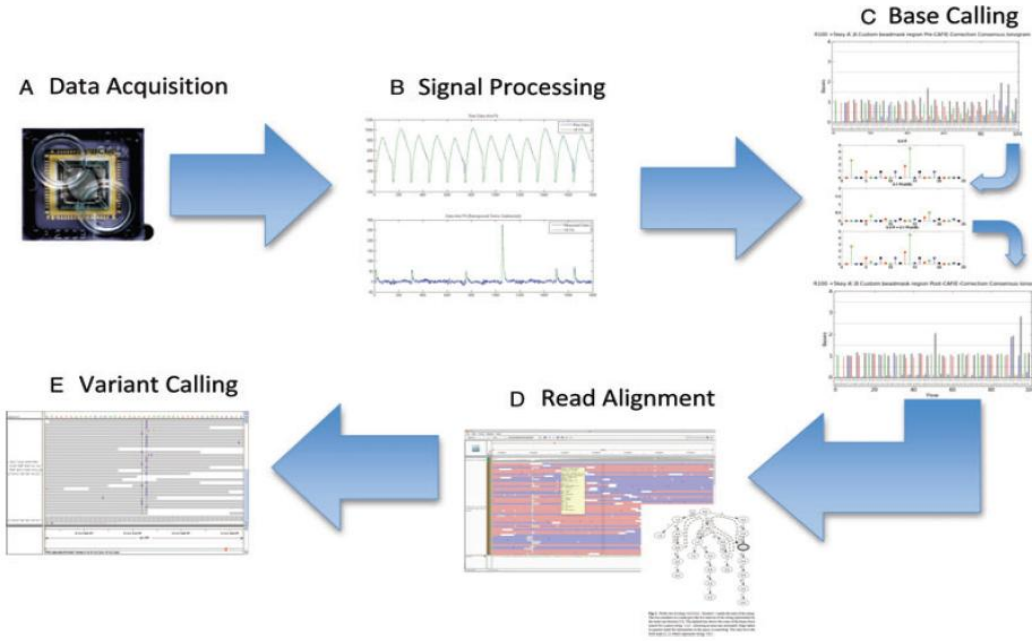

Figure 18. Overview of the data analysis process, which consists of: (A) data acquisition that requires high bandwidth data capture, transfer and reduction hardware solutions; (B) signal processing required for the removal of background pH variation and fit the incorporation model to the data; (C) base calling that consists of processing the sequencing signal to nucleotides; (D) alignment of the reads produced to a reference sequence and (D) variant calling that is when the differences between the samples and the reference genome are analyzed, and it is based on the coverage provided by the reads at each locus (taken from (124)).

After the base calling, the remainder of the bioinformatics analysis consists of mapping (Figure 18 – Read Alignment), variant calling and lastly annotation. The mapping step consists in the alignment of each short read to a position on a reference genome (in this case, the human genome), using software that will assess the likely starting point of each read within the reference genome. This step is hindered by the large volume of short reads generated by NGS, unique versus non-unique mapping and the variation in base quality. Beside this, the mapping is a critical step since any errors in this step will be carried through the rest of the analysis (126). The resulting sequence alignment is stored in a sequence alignment/map (SAM) or binary alignment/map (BAM) file (127). There are software available that differ in accuracy and speed that can be used to perform sequencing read alignment (126). In the case of Ion Proton, the platform provides the Torrent MAPper (TMAP) aligner to perform this step (124).

The next step in the pipeline is variant calling. This process usually refers to finding SNPs and small insertions/deletions (indels) and consists in the comparison of the aligned sequences to the reference genome to determine which positions deviate from the known position (126,128). These variants may have a functional impact, resulting in diseases or may simply be genetic variants with no functional effect (126). There are a few challenges regarding this step like the presence of indels, which represents a major source of false positive single nucleotide variant (SNV) identifications; errors from the library preparation due to PCR artifacts and variable GC content in the reads and finally, variable quality scores, being that most of the higher error rates are found at the end of reads (129). The variant call format (VCF) is the standard generic format where all the sequence variations are recorded (130). In Ion Proton, a specific variant calling software, the torrent variant caller (TVC) is provided (124).

The last step of the pipeline is annotation, the step in which all the known information about each variant that was detected is gathered. The annotation may reveal, among other things, if the variant is already known, if it has a functional impact or not and if that impact has already been predicted, if the function or activity of the specific gene is known and even if an identified disease is already associated with that variant in that gene. The result of this step is a small list of well-annotated variants that can explain a certain biological phenomenon (126). In this last step, different prediction programs are used to predict the effects of functional SNP. One of those is the polymorphism phenotyping v2 (PolyPhen-2) server, which is responsible for automated functional annotation of coding SNPs, predicting the damaging effects of genetic variants (131,132,133). For that, PolyPhen uses sequence conservation, structure and Universal Protein KnowledgeBase (UniProtKB) annotation (133). PolyPhen-2 possesses two databases, HumDiv (HDIV) and HumVar (HVAR). While HDIV should be used when accessing rare alleles at loci potentially involved in complex diseases, dense mapping of regions identified by GWAS and analysis of natural selection from sequence data, the HVAR should be used for diagnosis of Mendelian diseases, which requires distinguishing mutations with drastic effects from all the remaining human variation (132). Another prediction software is the sorting intolerant from tolerant (SIFT), which is based on the principles of protein evolution. This software is a multi-step algorithm that uses a sequence homology-based approach to classify amino acids substitutions, meaning that SIFT uses sequence homology to predict whether an

amino acid substitution will affect protein function and hence, potentially alter the phenotype (133,134). In this software, the scores are calculated using position-specific (hydrophobic conserved, highly conserved and unconserved) scoring matrices (133). Lastly, the other prediction tool used is the combined annotation dependent depletion (CADD). In this approach, CADD measures the deleteriousness of SNVs, a property that strongly correlates with both molecular functionality and pathogenicity, into a single score. This means that the basis of CADD is to quantitatively predict the deleteriousness, pathogenicity and molecular functionality, both protein-altering and regulatory, in a variety of experimental and disease contexts. This tool combines the generality of conservation-based metrics with the specificity of subset-relevant functional metrics (e.g. PolyPhen-2) (135).

7. Objectives

With the continued increase in the prevalence of T2D, the risk for the development of related complications, such as diabetic nephropathy, also increases. Currently, the underlying genetic mechanisms for the development of this complication remain unclear, and it is believed that SNPs in several genes distributed across the genome may be causal for the development and progression of the complication. Thus, the main objective of this work is to contribute to uncover the genetic mechanisms associated with the development of diabetic nephropathy. For that, this work intends to:

i. Identify common and rare genetic variants present in the exomes of 36 type 2 diabetic individuals with and without diabetic nephropathy using different statistical approaches;
ii. Search, in the results obtained for the 36 type 2 diabetic individuals, the presence of genes and genetic variants already associated with diabetic nephropathy identified based on the literature;
iii. Validate the biologically relevant genetic variants found in the studied population;
iv. Establish a phenotype-genotype correlation between the genetic variants present in the exomes of the individuals in the study and the presence or absence of diabetic nephropathy.

This study will, therefore, identify the relevant common and rare genetic variants and genes associated with diabetic nephropathy in Portuguese type 2 diabetic patients, and highlight if there is a particular mechanism involved in disease pathophysiology within the Portuguese population. The genetic variants and genes uncovered in this study can be used in following studies to contribute to a deeper understanding of the genetic mechanisms associated with diabetic nephropathy.

CHAPTER 2| Materials and Methods

1. Characterization of the population under study

This study was performed in a 36 Portuguese patients group, comprising 20 men and 16 women between the ages of 45 and 77 years old with diagnosed T2D. These patients were divided, based on their phenotypes, into a control group formed by 19 individuals with T2D and without diabetic nephropathy and a case group consisting of 17 individuals with T2D that presented the diabetic complication. The peripheral blood samples collected from the diabetic individuals that participated in this study were obtained from the Endocrinology Unit of the Coimbra Hospital and University Center. The diabetic nephropathy diagnosis on the subjects was made by a physician from the same institution. Informed written consent was obtained from all participants before the study.

Case and control groups were characterized by age, disease duration, glycated hemoglobin (HbA1c) levels and sex, using the International Business Machines Corporation (IBM) Statistical Package for the Social Sciences (SPSS) software, version 20.0 (International Business Machines Corporation, Armonk, New York, United States of America). For the qualitative covariate, sex, a chi-square (χ^2) test was performed while, for the quantitative covariates, age, disease duration, and HbA1c, a t-test was conducted.

2. Whole-Exome Sequencing

2.1. DNA extraction and Quality Control

The DNA extraction process from the peripheral blood samples was performed using an extraction kit, the DNeasy Blood & Tissue Kit (QIAGEN®, Hilden, Germany). The workflow of this kit is presented in Figure 19.

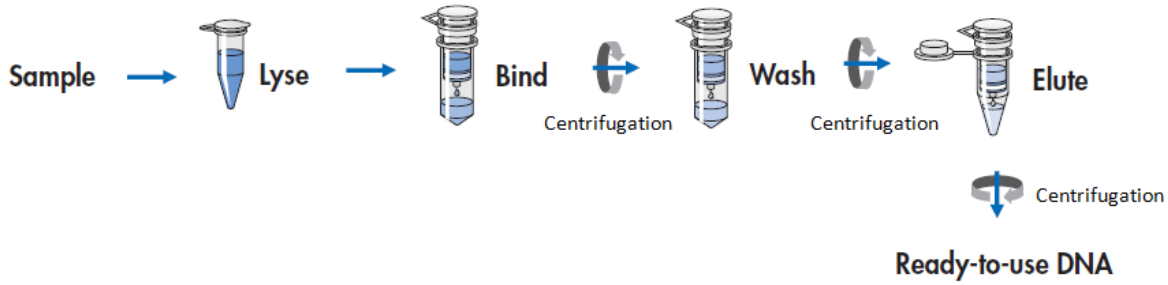

Figure 19. Workflow of the DNA extraction process using the DNeasy Blood & Tissue Kit (adapted from (136)). DNA: deoxyribonucleic acid.

The DNA samples are first lysed using a lysis buffer containing a detergent, which is responsible for the disruption of the cell and nuclear membranes, and proteinase K, to digest the protein components. Subsequently, ethanol is added to precipitate the DNA and the purification process begins. The lysate is loaded onto a DNeasy Mini spin column and then centrifuged, causing the DNA to bind selectively to the DNeasy membrane and the contaminants to pass through. The buffering conditions are adjusted to provide optimal DNA binding conditions. The remaining contaminants and enzyme inhibitors are then removed in two wash steps and the DNA is eluted in 100 microliters (μL) of the elution buffer.

Following the extraction, the DNA was quantified, and its purity was assessed by spectrophotometry, applying two μL of each DNA sample to the NanoDrop® ND-1000 Spectrophotometer (NanoDrop Technologies, Wilmington, United States of America). This technology allows the determination of nucleic acids concentration, both single stranded as well as double stranded, based on the Beer-Lambert law ([DNA] (nanograms per microliter (ng/μL)) = A_{260} x 50 micrograms per milliliter (μg/mL) x dilution factor, where A_{260} represents the absorbance at 260 nm, the maximum peak of absorbance for nucleic acids). The purity of the samples or the presence of contaminants such as proteins, phenols, and RNA, among others, was determined by the ratios 260nm/280nm and 260nm/230nm. A 260nm/280nm ratio between 1.7 and 1.9 is generally accepted as pure DNA, while in the 260nm/230nm ratio, expected values are commonly in the range of 2.0-2.2. The samples that had contamination (ratios lower than the referred values) were subjected to a standard purification with isopropanol.

After the purifying steps, the integrity of the samples was evaluated by performing an electrophoresis on a 1% (weight/volume (w/v)) agarose gel with tris-acetate-ethylenediamine tetraacetic acid (TAE) buffer, in which 50 nanograms (ng) of DNA were run along with the molecular marker NZYDNA Ladder III (NZYTech, Lisbon, Portugal). A current of 90 volts (V) for 30 minutes was applied and the gel, stained with ethidium bromide, was visualized under ultraviolet light with the Molecular Imager Gel Doc XR System (Bio-Rad Laboratories, Hercules, California, United States of America).

Finally, another quantification protocol was performed, using for the sample preparation the Qubit® dsDNA HS Assay Kit (Invitrogen™, Life Technologies™, Eugene, Oregon, United States of America) and for the quantification itself, the Qubit® 2.0 Fluorometer

(Invitrogen™, Life Technologies™, Carlsbad, California, United States of America). This technology uses fluorescent probes that connect specifically to DNA to quantify the sample. A more precise value of DNA concentration is obtained than the one obtained with NanoDrop® and without the possible contaminants. A DNA sample concentration higher than 10 ng/µL was considered acceptable.

All of the procedures described above, as well as the kits used, were performed accordingly to the manufacturer instructions.

2.2. Sequencing process

The first step in the sequencing process consists of an exome library preparation for each DNA sample. The Ion Ampliseq™ Library Kit 2.0 (Life Technologies™, Carlsbad, California, United States of America) was used. For each sample, a multiplex PCR reaction with 294,000 pairs of primers distributed across 12 pools of primers allowed the amplification of the genome coding regions (exome), being this step designed to create overlapping amplicons, to cover all the target regions. The kit master mix was distributed across 12 wells, followed by 75 ng of DNA from a single sample per well (Figure 20).

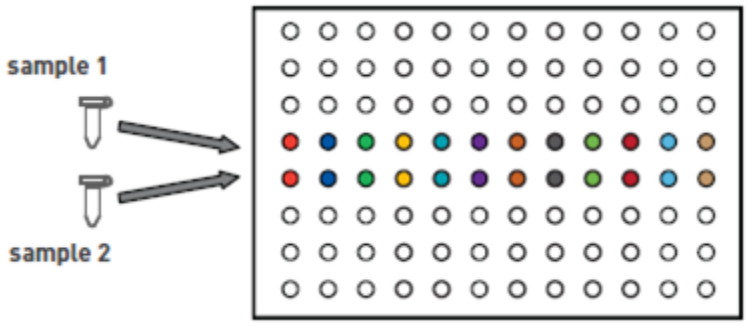

Figure 20. Example of the plate used in the library preparation for exome sequencing (adapted from (137)).

After this process, the samples underwent a multiplex PCR in which all the exons were amplified, followed by the combination of the amplicons by sample into a single plate well, where FuPa reagent was added to digest partially primer sequences. Subsequently, each DNA fragment was ligated to adapters with barcodes using the Ion Xpress™ Barcode Adapters 1-16 Kit (Life Technologies™, Carlsbad, California, United States of America), allowing the distinction between fragments from two different DNA samples. This permits

the sequencing of different DNA samples at the same time (Figure 21). Lastly, the exome library preparation ends with the purification of the amplified library by adding magnetic beads, the Agencourt® AMPure® XP (Beckman Coulter Inc., Brea, California, United States of America), that connect to the DNA fragments with the adapters and elute the unamplified fragments as well as the digested primers.

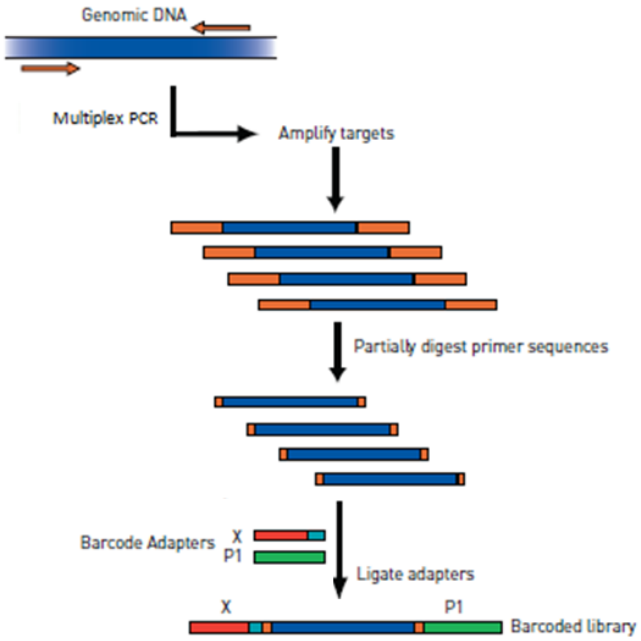

Figure 21. Exome library preparation using the Ion AmpliseqTM Library Kit 2.0 (adapted from (137)). DNA: deoxyribonucleic acid; PCR: polymerase chain reaction.

The next step was to assess the exome library profile through evaluation of the DNA fragment size and to verify if primer-dimers were present. The library size profile was determined with the High Sensitivity DNA Kit (Agilent Technologies®, Waldbronn, Germany) in the Agilent 2100 Bioanalyzer and the software 2100 expert (Agilent Technologies®, Santa Clara, California, United States of America). The concentration of each exome library was determined by real-time PCR with the Ion Library TaqMan® Quantitation Kit (Applied Biosystems®, Life Technologies™, Austin, Texas, United States of America) in a 7500 Fast Real-Time PCR System and the results analyzed with the 7500 Software v2.0 (Applied Biosystems®, Life Technologies™, Singapore). Proper dilutions that resulted in an exome library concentration of approximately 100 picomolar (pM) were prepared for template amplification and exome sequencing on the Ion Proton platform.

The following stage in the sequencing process is the clonal amplification of each DNA fragment by ePCRs, to generate enough copies of the sequencing template. In the ePCR first step, the adapters connected to the DNA fragment are matched with the complementary oligonucleotides connected to the Ion Sphere™ particles (ISPs). After, each ISP connected to the DNA fragment is isolated in an individual water-in-oil emulsion containing the necessary reagents for the amplification process to occur. Each ePCR combined two exome libraries. Following the ePCR, the water-in-oil emulsions are broken allowing the recovery of the ISPs. This entire procedure was carried out on the Ion OneTouch™ 2 System (Life Technologies™, Carlsbad, California, United States of America) using the Ion PI™ Template OT2 200 Kit v2 (Life Technologies™, Carlsbad, California, United States of America).

After the ISPs recovery, an aliquot was saved to perform a quality control test. Each ISP contains multiple copies of the same DNA fragment, being those fragments bound to the ISP on one side and biotinylated on the other side. During enrichment, the biotinylated side is bound to magnetic beads that are coated with streptavidin, allowing to retain the template-positive ISPs (ISPs that have amplified DNA) and wash away all empty ISPs that do not contain any DNA fragment. After the enrichment, another aliquot was saved. The quality control test was performed with both aliquots, before and after enrichment, to determine the percentage of enriched ISPs in the total ISPs. This test was carried out on the Qubit® 2.0 Fluorometer (Invitrogen™, Life Technologies™, Carlsbad, California, United States of America) using the Ion Sphere™ Quality Control Kit (Life Technologies™, Carlsbad, California, United States of America). Lastly, a chip was loaded with the enriched ISPs using the Ion PI™ Chip Kit v2 (Life Technologies™, Carlsbad, California, United States of America) as well as the Ion PI™ Sequencing 200 Kit v2 (Life Technologies™, Carlsbad, California, United States of America), and the sequencing process was executed in the Ion Proton™ Sequencer (Life Technologies™, Carlsbad, California, United States of America). Each chip was loaded with two exome libraries, the process being repeated for 18 sequencing runs in a total of 36 DNA samples/exomes.

All of the procedures, as well as the kits described above, were performed accordingly to the manufacturer's instructions.

2.3. Bioinformatics Analysis

Following the completion of a sequencing run, the Torrent Suite™ Software (Life Technologies™, Carlsbad, California, United States of America) processed the raw data obtained. This software contains all the necessary software components to execute signal processing, base calling, read alignment (mapping) and variant calling (Chapter 1| Introduction – Figure 18) and is already installed on the Proton™ Torrent Server (Life Technologies™, Carlsbad, California, United States of America).

For each of the 36 sequenced exomes, the adapter sequences, as well as the low-quality bases were trimmed using the Torrent Suite™ Software (Life Technologies™, Carlsbad, California, United States of America). The resulting reads were then mapped to the human reference genome hg19, using the TMAP version 4.0.6 (Life Technologies™, Carlsbad, California, United States of America), and the aligned sequence stored in a BAM file.

For a sample identification control, the number of reads mapped to the Y chromosome were counted using SAMtools, a library, and software package for manipulating alignments in SAM or BAM files (127), and divided by the number of total mapped reads thereby creating a ratio to determine sample sex. Therefore, this test determined, based on the sequencing results, the sex of each sequenced exome, and compared this "informatics determined sex" to the sex of each individual in the study.

The next step in this pipeline is the variant calling, a step where the positions in which the aligned sequence deviates from the reference sequence, creating genetic variants, are identified. This step was performed by running the TVC plugin version 4.0 (Life Technologies™, Carlsbad, California, United States of America) with the optimized parameters for exome sequencing recommended for the Ion AmpliSeq™ Exome Kit (Life Technologies™, Carlsbad, California, United States of America), being the genetic variants recorded into a VCF file. The genotypes were then recalled based on the altered allele frequency, using for that an in-house script. Therefore, altered homozygous genotypes were assigned for altered allele frequencies higher than 0.8 and the heterozygous genotypes were assigned for frequencies between 0.2 and 0.8. The homozygous positions equal to the reference genome were further evaluated using an in-house script since no information about these genotypes is reported in the generated VCF. The in-house script stipulated that when the altered allele frequencies were lower than

0.05, the genotype was considered homozygous equal to the reference, but, if the altered allele frequencies were higher than 0.05, then the genotype was considered undetermined. Furthermore, the genetic variants with less than 10x depth of coverage were discarded. The remaining variants from the 36 exomes were considered and integrated altogether using VCFtools, an open-source software package for analyzing and manipulating VCF files (130).

The last step in the bioinformatics analysis is the annotation. This is the step in which all the known information about each genetic variant detected in the 36 exomes is gathered. For this process, the Variant Effect Predictor (VEP) software, responsible for predicting the impact of genetic variants in genes, transcripts and protein sequences (138), was used, being the transcripts used by this software obtained from the version GRCh37 from Ensembl (139). The VEP results were then used by the Genome Mining (GEMINI) framework (140) to annotate the genetic variants against a comprehensive set of genomic annotation files, which includes the Single Nucleotide Polymorphism Database (dbSNP) (141), the Encyclopedia of DNA Elements (ENCODE) (142), the Clinical Variation database (ClinVar) (143), the 1000 Genomes Project (1000G) (144), the Exome Sequencing Project (ESP) (145), the Kyoto Encyclopedia of Genes and Genomes (KEGG) (146), the Genomic Evolutionary Rate Profiling (GERP) (147), and lastly the Human Protein Reference Database (HPRD) (148). Finally, GEMINI saved all information from the impact and annotation prediction in a database for further analysis. Furthermore, to detect relatedness between samples, pair-wise genetic distance tests were performed by the GEMINI framework.

3. Common Variants Approach

3.1. Statistical Analysis

The common variants approach was used to test for an association between diabetic nephropathy and several single common genetic variants. It consisted in the prioritization of the common variants using the χ^2 test or the Fisher exact test to detect the ones that had significant genotype differences between cases and controls (univariate analysis). The analysis was performed in the IBM SPSS software, version 20.0 (International Business

Machines Corporation, Armonk, New York, United States of America). Only variants that passed the Hardy-Weinberg equilibrium (HWE) (p-value > 0.05) were considered. From the obtained genetic variants, the ones with a p-value ≤ 0.05 and a call rate ≥ 0.9 were then tested in a binary logistic regression model for adjustment to factors like age, disease duration, HbA1c, and sex, being statistically significant results considered for a p-value ≤ 0.05. As the number of statistically significant genetic variants obtained was high, several filters were applied to prioritize these variants. Therefore, the variants present in the 3' or 5'-untranslated regions (UTR), those intronic, the ones upstream or downstream of the gene, as well as the synonymous variants, had to have a CADD (135) > 15 and be in conserved regions. Missense variants had to have at least one of the prediction programs (PolyPhen-2 (132), SIFT (134), and CADD) indicating that the variant was pathogenic or the variant had to be in a conserved region.

Moreover, each one of the common variants that passed the filters was studied for its gene function and expression, as well as for any known disease association, to detect the most biologically relevant to the pathogenesis of diabetic nephropathy. The gene function was evaluated through GeneCards®: The Human Gene Database (www.genecards.org) (149) and HumanMine (www.humanmine.org) (150), the gene expression was evaluated through neXtProt (www.nextprot.org) (151) and also HumanMine, and the variants association to any known disease was studied based on the information contained in the Human Gene Mutation Database (HGMD) (www.hgmd.cf.ac.uk) (152) and ClinVar (www.ncbi.nlm.nih.gov/clinvar) (143). Furthermore, Google (www.google.pt) and PubMed (www.ncbi.nlm.nih.gov/pubmed) were also used to perform a more generalized search for the genes in which the genetic variants were present and to search for the variants.

Lastly, for the biologically relevant genetic variants found, its functional impact was assessed based on information provided by VEP and then used by the GEMINI framework, as well as information obtained from UniProtKB (www.uniprot.org) (153).

3.2. Candidate Genes

In order to compile a list of candidate genes and respective common variants already associated with diabetic nephropathy from previous studies, a search was made using

Google (www.google.pt), PubMed (www.ncbi.nlm.nih.gov/pubmed), HGMD (www.hgmd.cf.ac.uk) (152) and ClinVar (www.ncbi.nlm.nih.gov/clinvar) (143). The research was restricted to candidate genes and their common variants associated with diabetic nephropathy only in European type 2 diabetic individuals, being this list then compared with the results obtained from the statistical analysis of common variants in the 36 exomes used in this study.

3.3. Validation

Several methods were used to validate the common variants found in the 36 exomes studied. The first validation method was performed with the objective of verifying if the variants found were real or simply an error of the sequencing technology. For that, the BAM file containing the aligned sequence of each exome was manually verified for the positions of the genetic variants present in that same exome. The second validation method consisted in genotyping 4 of the 36 exomes used in this study by Illumina microarray with the HumanOmniExpressExome (Illumina®, San Diego, California, United States of America) at ATLAS Biolabs GmbH in Berlin, Germany. Besides the 4 of the 36 exomes in study, others exomes also sequenced by the Ion Proton™ Sequencer (Life Technologies™, Carlsbad, California, United States of America) were genotyped by this method, which allowed the validation of the results obtained by this sequencing technology.

4. Rare Variants Approach

4.1. Statistical Analysis

The rare variants approach was used to test for an accumulation of rare variants in genes by a gene-wise burden test. GEMINI (140) was used to select the rare variants that passed the HWE (p-value > 0.05) and those with a call rate ≥ 0.9, a MAF ≤ 0.01 in 1000G (144) and/or ESP (145) and a MAF ≤ 0.05 in the study population. Variants had to be non-synonymous or present in splice site regions. The rare variants that passed the filters were then prioritized through a binary (cases vs. controls) gene-wise burden test, the emmaxVT test, using the Efficient and Parallelizable Association Container Toolbox (EPACTS) (154). Statistically significant results were considered for a p-value ≤ 0.05.

From the resulting list, and in a similar manner to the common variants approach, each one of the statistically significant genes with accumulated rare variants was studied for its function and expression, as well as for any known disease association from the rare variants present in that gene, in order to detect the most biologically relevant genes to the development of diabetic nephropathy. The gene function was evaluated through GeneCards®: The Human Gene Database (www.genecards.org) (149) and HumanMine (www.humanmine.org) (150), the gene expression was evaluated through neXtProt (www.nextprot.org) (151) and also HumanMine, and the association of the rare variants present in a particular gene to any known disease was studied based on the information contained in the HGMD (www.hgmd.cf.ac.uk) (152) and ClinVar (www.ncbi.nlm.nih.gov/clinvar) (143). Furthermore, to perform a more generalized search for the genes, Google (www.google.pt) and PubMed (www.ncbi.nlm.nih.gov/pubmed) were also used.

Lastly, for the genes considered as biologically relevant, their accumulated rare variants functional impact was assessed based on information provided by VEP and then used by the GEMINI framework, as well as information obtained from UniProtKB (www.uniprot.org) (153).

4.2. Validation

To validate the genes obtained by the statistical analysis, several validation methods were used. The first method, as it was for the common variants, was performed with the objective of verifying if the rare variants accumulated in the genes found were real or an error of the sequencing technology, using for that the BAM file of each exome to manually verify the positions of the rare genetic variants present in that same exome. The second validation method consisted in the determination of each rare variant MAF in the sequenced exomes of 19 healthy (without T2D and without diabetic nephropathy) individuals unrelated to this study. Thirdly, some rare variants, like what happen with the common variants, were validated through genotyping 4 of the 36 exomes used in this study by the Illumina microarray with the HumanOmniExpressExome (Illumina®, San Diego, California, United States of America). This validation method was performed by ATLAS Biolabs GmbH in Berlin, Germany.

The rare variants that were not validated by this latter method were subjected to allele-specific oligonucleotide polymerase chain reaction (ASO-PCR). To test the presence of a SNP by this method, each primer, either forward or reverse, was designed to be complementary and specific for only one allele, one primer was specific for the reference allele and the other primer was specific for the altered allele of the DNA being tested, having both primers to have the same direction (both forward or both reverse). The remaining primer is a common one to both the allele-specific primers. Furthermore, a pair of control primers was also added in each PCR reaction to ensure that all amplification conditions were met. The primers were designed in the Oligo Explorer software, version 1.2 (Gene Link™, Hawthorne, New York, United States of America) and are listed in Table 5. For each ASO-PCR reaction, the conditions had to be optimized using BIOTAQ™ DNA Polymerase (Bioline, London, United Kingdom), deoxynucleotides (dNTPs) (BIORON, Ludwigshafen, Germany) and, for some of the rare variants, betaine solution (VWR, Denmark), being the reaction final volume 10.1 μL. The used reagents conditions are presented in Table 6. The amplification process was performed by MyCycler™ Thermal Cycler (Bio-Rad Laboratories, Hercules, California, United States of America) under the conditions presented in Table 7. Lastly, the results were visualized by performing an electrophoresis on a 1.5% (w/v) agarose gel with TAE buffer, in which, besides the samples, a molecular marker, the NZYDNA Ladder I (NZYTech, Lisbon, Portugal). A current of 90 V was applied for 40 minutes to the agarose gel. The gel was stained with ethidium bromide and visualized under ultraviolet light with the Molecular Imager Gel Doc XR System (Bio-Rad Laboratories, Hercules, California, United States of America). Negative controls were performed, along with a wild-type (homozygous equal to the reference) and a heterozygous sample, for each reaction, using the same conditions.

Table 5. Information regarding the primers used for validation by ASO-PCR.

Gene	rs ID or Chr: end position	Primer name	Primer sequence (5'>>3')	Frag. size (bp)	Control primer name	Control primer sequence (5'>>3')	Control frag. size (bp)	Optimized Ta (ºC)
STAB1	rs371042844	STAB1_NF_F_C STAB1_NF_F_T STAB1_NF_R	CCATCCTGGAGGTAAGCTC CCATCCTGGAGGTAAGCTT CCACAGCCTCATTGCTTGG	370	ITGA1_DR_F ITGA1_DR_R	CTCAGGAGAAAGCAGTTGTG CACCCATCCAACATGAAGAC	537	66
	rs41292856	STAB1_NF1_F_N STAB1_NF1_R_A_N STAB1_NF1_R_G_N	CTTCTCCATCTTCTACCAATG GTTCACCACATACCCCACT GTTCACCACATACCCCACC	639	PRKCQ_DR_F PRKCQ_DR_R	GTGCTCTGTCCTCCTTATAC GGTATTCGTCTTGGCATCTC	355	64
	rs149944392	STAB1_NF2_F STAB1_NF2_R_G STAB1_NF2_R_A	GCTTCAGTGATCTGAGTCAG GCCATCCCCGCTGTAACC GCCATCCCCGCTGTAACT	433	ITGA1_DR_F ITGA1_DR_R	CTCAGGAGAAAGCAGTTGTG CACCCATCCAACATGAAGAC	537	64
	rs199636230	STAB1_NF3_F_C STAB1_NF3_F_T STAB1_NF3_R	CAACCATGACTCCACTTGCTC CAACCATGACTCCACTTGCTT CTCCAGCCCACAGATGCTG	488	PRKCQ_DR_F PRKCQ_DR_R	GTGCTCTGTCCTCCTTATAC GGTATTCGTCTTGGCATCTC	355	67
	rs143836348	STAB1_NF4_F STAB1_NF4_R_T STAB1_NF4_R_C	CTTCTCTGCCTTCCAGGTAG GCCACACTTCTTCACTTCCA GCCACACTTCTTCACTTCCG	239	ITGA1_DR_F ITGA1_DR_R	CTCAGGAGAAAGCAGTTGTG CACCCATCCAACATGAAGAC	537	66
	Chr 3: 52546850	STAB1_NF5_F_N STAB1_NF5_R_G_N STAB1_NF5_R_A_N	CCACTTCTCCATCTTCTACC GGCAGGAAGCGTGATGCC GGCAGGAAGCGTGATGCT	239	ITGA1_DR_F ITGA1_DR_R	CTCAGGAGAAAGCAGTTGTG CACCCATCCAACATGAAGAC	537	66
	Chr 3: 52548167	STAB1_NF6_F STAB1_NF6_R_G STAB1_NF6_R_A	CATCCTCAGCCAGGTACAG GGCACAAAGATGGTGTAGGC GGCACAAAGATGGTGTAGGT	389	IOP1_DR_F IOP1_DR_R	GTGAAGGGTGAAGGTCAGTG CAGGTGAATTCGACCAGTGG	532	69

Gene	rs ID or Chr: end position	Primer name	Primer sequence (5'>>3')	Frag. size (bp)	Control primer name	Control primer sequence (5'>>3')	Control frag. size (bp)	Optimized Ta (ºC)
MMP25	Chr 16: 3108251	MMP25_NF_F_T MMP25_NF_F_C MMP25_NF_R	CGCACGGCTGCACCGCTT CGCACGGCTGCACCGCTC GTCACCTGGAGGAGCAGAG	599	PRKCQ_DR_F PRKCQ_DR_R	GTGCTCTGTCCTCCTTATAC GGTATTCGTCTTGGCATCTC	355	-
		MMP25_NF_F_T MMP25_NF_F_C MMP25_NF_R2_N	CGCACGGCTGCACCGCTT CGCACGGCTGCACCGCTC CTCGCACTGACAATCGCAGG	782	IOP1_DR_F IOP1_DR_R	GTGAAGGGTGAAGGTCAGTG CAGGTGAATTCGACCAGTGG	532	
		MMP25_NF_F_N MMP25_NF_R_T_N MMP25_NF_R_C_N	GTGTCAGCCTCCCAATGTGC CGGGCAGCCCCTCCCAGA CGGGCAGCCCCTCCCAGG	318	IOP1_DR_F IOP1_DR_R	GTGAAGGGTGAAGGTCAGTG CAGGTGAATTCGACCAGTGG	532	
	Chr 16: 3108573	MMP25_NF1_F_C MMP25_NF1_F_A MMP25_NF_R	GCAGTACTGGCGCTACGAC GCAGTACTGGCGCTACGAA GTCACCTGGAGGAGCAGAG	278	IOP1_DR_F IOP1_DR_R	GTGAAGGGTGAAGGTCAGTG CAGGTGAATTCGACCAGTGG	532	66
CUX1	Chr 7: 101758496	CUX1_NF_F_A CUX1_NF_F_T CUX1_NF_R	GCCTTTCAGCCCTGGAAAA GCCTTTCAGCCCTGGAAAT CTAGCGTTGTTAGGTGTGAC	507	PRKCQ_DR_F PRKCQ_DR_R	GTGCTCTGTCCTCCTTATAC GGTATTCGTCTTGGCATCTC	355	62

bp: base-pairs; ºC: degrees Celsius; Chr: chromosome; Frag.: fragment; Ta: annealing temperature.

Table 6. Final concentration of the reagents used in ASO-PCR.

Rare Variants / Reagents	rs371042844	rs41292856	rs149944392	rs199636230	rs143836348	Chr 3: 52546850	Chr 3: 52548167	Chr 16: 3108251	Chr 16: 3108573	Chr 7: 101758496
10x NH_4 Reaction Buffer	3.33x	3.33x	3.33x	3.33x	3.33x	3.33x	3.33x	3.33x	3.33x	3.33x
$MgCl_2$ Solution (50 mM)	7.08 mM	7.08 mM	7.08 mM	7.08 mM	7.08 mM	7.08 mM	7.08 mM	7.08 mM	7.08 mM	7.08 mM
dATP (100 mM)	0.67 mM	0.67 mM	0.67 mM	0.67 mM	0.67 mM	0.67 mM	0.67 mM	0.67 mM	0.67 mM	0.67 mM
dCTP (100 mM)	0.67 mM	0.67 mM	0.67 mM	0.67 mM	0.67 mM	0.67 mM	0.67 mM	0.67 mM	0.67 mM	0.67 mM
dGTP (100 mM)	0.67 mM	0.67 mM	0.67 mM	0.67 mM	0.67 mM	0.67 mM	0.67 mM	0.67 mM	0.67 mM	0.67 mM
dTTP (100 mM)	0.67 mM	0.67 mM	0.67 mM	0.67 mM	0.67 mM	0.67 mM	0.67 mM	0.67 mM	0.67 mM	0.67 mM
Betaine Solution (5 M)	-	-	-	-	-	-	-	1.50 M	1.50 M	-
Ref. allele primer (10 µM)	2.00 µM	2.00 µM	2.00 µM	2.00 µM	2.00 µM	2.00 µM	1.00 µM	2.00 µM	2.00 µM	2.00 µM
Alt. allele primer (10 µM)	2.00 µM	2.00 µM	2.00 µM	2.00 µM	2.00 µM	2.00 µM	1.00 µM	2.00 µM	2.00 µM	2.00 µM
Common primer (10 µM)	2.00 µM	2.00 µM	2.00 µM	2.00 µM	2.00 µM	2.00 µM	1.00 µM	2.00 µM	2.00 µM	2.00 µM
Control forward primer (10 µM)	0.50 µM	1.00 µM	0.50 µM	1.50 µM	2.00 µM	1.00 µM	0.75 µM	2.00 µM	0.50 µM	1.50 µM
Control reverse primer (10 µM)	0.50 µM	1.00 µM	0.50 µM	1.50 µM	2.00 µM	1.00 µM	0.75 µM	2.00 µM	0.50 µM	1.50 µM
BIOTAQ DNA Polymerase (5 U/µL)	0.05 U/µL	0.05 U/µL	0.05 U/µL	0.05 U/µL	0.05 U/µL	0.05 U/µL	0.05 U/µL	0.05 U/µL	0.05 U/µL	0.05 U/µL
DNA (20 ng/µL)	3.96 ng/µL	3.96 ng/µL	3.96 ng/µL	3.96 ng/µL	3.96 ng/µL	3.96 ng/µL	3.96 ng/µL	3.96 ng/µL	3.96 ng/µL	3.96 ng/µL

Alt.: altered; Chr: chromosome; dATP: deoxyadenosine triphosphate; dCTP: deoxycitidine triphosphate; dGTP: deoxyguanosine triphosphate; DNA: deoxyribonucleic acid; dTTP: deoxythymidine triphosphate; M: molar; $MgCl_2$: magnesium chloride; µM: micromolar; mM: millimolar; ng/µL: nanograms per microliter; NH_4: ammonium; Ref.: reference; U/µL: units per microliter.

Table 7. Amplification conditions used in ASO-PCR.

Temperature (ºC)	Time	Cycles
94	3 minutes	1
94	30 seconds	30
Optimized Ta for each variant	45 seconds	
72	1 minute	
72	2 minutes	1

ºC: degrees Celsius; Ta: annealing temperature.

Sanger sequencing was also used in the validation process for the rare variants approach. However, this method was only used as a confirmation method, since it was performed to confirm the genotype of a single exome in which doubts had arisen in its validation by the previously mentioned methods. This sequencing method is based on the ability of DNA polymerases to incorporate dNTPs and dideoxynucleotides (ddNTPs). The DNA polymerases copy single-stranded DNA templates by adding dNTPs to a growing chain, however, when they incorporate ddNTPs at the 3' end of the growing chain, chain elongation is prematurely terminated on a specific nucleotide (A, T, C or G), resulting in different size fragments. Each of the chain-terminating ddNTPs are differently fluorescently labeled, which, along with the size difference of each fragment, makes it possible to separate the fragments by capillary electrophoresis and "construct" the DNA sequence. Therefore, for the genotype confirmation by Sanger sequencing, the gene region of interest was amplified using a forward and reverse primer, designed in the Oligo Explorer software, version 1.2 (Gene Link^{TM}, Hawthorne, New York, United States of America) (Table 8).

Table 8. Information regarding the primers used for the amplification of the region of interest for Sanger sequencing.

Gene	Chr: end position	Primer name	Primer sequence (5'>>3')	Frag. size (bp)	Optimized Ta (ºC)
MMP25	Chr 16: 3108251	MMP25_NF_F_N MMP25_NF_R	GTGTCAGCCTCCCAATGTGC GTCACCTGGAGGAGCAGAG	882	62

bp: base-pairs; ºC: degrees Celsius; Chr: chromosome; Frag.: fragment; Ta: annealing temperature.

For this PCR reaction, the conditions were optimized using Phusion High-Fidelity DNA Polymerase (Thermo Scientific, Lithuania), dNTPs (BIORON, Ludwigshafen, Germany) and betaine solution (VWR, Denmark), with a final reaction volume of 50 μL. The reagents final concentrations are presented in Table 9.

Table 9. Final concentration of the reagents used in the PCR reaction.

Rare Variant / Reagents	Chr 16: 3108251
5x Phusion GC Buffer	1.00x
dNTPs mix (5 mM)	0.20 mM
Betaine (5 M)	1.50 M
Forward primer (10 μM)	0.30 μM
Reverse primer (10 μM)	0.30 μM
Phusion DNA Polymerase (2 U/μL)	0.02 U/μL
DNA (20 ng/μL)	2.00 ng/μL

Chr: chromosome; DNA: deoxyribonucleic acid; dNTPs: deoxynucleotides; M: molar; μM: micromolar; mM: millimolar; ng/μL: nanograms per microliter; U/μL: units per microliter.

The amplification process was performed by MyCycler^{TM} Thermal Cycler (Bio-Rad Laboratories, Hercules, California, United States of America) under the conditions presented in Table 10. The region of interest in this gene had a high percentage of GC content, and a touchdown polymerase chain reaction (TD-PCR) was selected for its amplification. This type of PCR represents an empirical approach to favoring the most specific primer–template interactions, and can be incorporated as a standard part of any PCR to enhance specificity and product formation. The key principle in TD-PCR is to employ successively lower annealing temperatures, beginning with an annealing temperature above the projected melting temperature, then transitioning to a lower, more permissive temperature over the course of 10–15 cycles. This approach exploits the exponential nature of PCR, where the first stages of annealing and amplification are the most critical in producing the desired product (155).

Table 10. Amplification conditions used in TD-PCR.

Temperature (ºC)	Time	Cycles
98	3 minutes	1
99	1 minute	1
98	10 seconds	14
76-1/cycle	30 seconds	
72	30 seconds	
98	10 seconds	21
62	30 seconds	
72	30 seconds	
72	5 minutes	1

ºC: degrees Celsius.

Lastly, the results were visualized by performing an electrophoresis on a 1% (w/v) agarose gel with TAE buffer, in which, besides the sample, a molecular marker, the NZYDNA Ladder I (NZYTech, Lisbon, Portugal). A current of 90 V was applied for 30 minutes to the agarose gel and the gel, stained with ethidium bromide, was visualized under ultraviolet light with the Molecular Imager Gel Doc XR System (Bio-Rad Laboratories, Hercules, California, United States of America). A negative control, as well as a wild-type and a heterozygous sample, was performed for this PCR reaction, using the same conditions.

After this process, the amplified fragment was purified using magnetic beads, the Agencourt® AMPure® XP (Beckman Coulter Inc., Brea, California, United States of America), that connect to the amplified DNA fragment, allowing the unamplified material to be discarded. This procedure was performed accordingly to the manufacturer instructions. Finally, the purified fragment was sent to GATC Biotech in Cologne, Germany, for Sanger sequencing.

CHAPTER 3| Results and Discussion

1. Characterization of the population under study

Case-control studies are used to compare the frequency of SNP alleles in two different but well-defined groups of individuals, the cases, who have been diagnosed with the disease under study, and the controls, who are either known to be unaffected or who have been randomly selected from the population. One of the major problems in this type of studies is ensuring a good match between the genetic background of cases and controls so that any genetic difference between them is related to the disease under study and not to biased sampling. Therefore, cases and controls should be from similar ethnic groups and have similar characteristics (156).

A total of 36 individuals were included in this study, 20 (55.6%) men and 16 (44.4%) women. From those individuals, 19 subjects composed the control group and 17 subjects were included in the case group, according to the nephropathy clinical diagnosis. The characteristics of each exome used in this study are presented in Appendix B – Table B1 and Table B2, while the overall characteristics of both groups are presented in Table 11.

Table 11. Statistics of the individuals in the control and case groups for the covariates used in the statistical analysis.

Covariates \ Statistics			Mean ± Standard Deviation	t-test (p-value)	%	χ^2 (p-value)
Age (years)	**Controls (n=19)**		62.2 ± 8.1	0.022	-	-
	Cases (n=17)		67.8 ± 5.7			
Disease duration (years)	**Controls (n=19)**		14.7 ± 9.5	0.556	-	-
	Cases (n=17)		16.5 ± 9.1			
HbA1c (%)	**Controls (n=19)**		9.6 ± 2.2	0.279	-	-
	Cases (n=17)		8.7 ± 2.6			
Sex	**Controls (n=19)**	**Male (n=9)**	-	-	47.4	0.296
		Female (n=10)			52.6	
	Cases (n=17)	**Male (n=11)**	-	-	64.7	
		Female (n=6)			35.3	

χ^2: chi-square test; HbA1c: glycated hemoglobin; n: number of individuals.

The covariate age was the only one that presented a statistically significant difference between the control and case groups (p-value ≤ 0.05), while the remaining covariates disease duration, HbA1c and sex, showed no statistically significant difference between the two groups (p-value ≥ 0.05). However, despite the difference between the groups for age, the performed statistical tests were adjusted for this parameter, as well as for the remaining covariates.

One of the major determinants of the development of microalbuminuria or the progression of microalbuminuria (incipient diabetic nephropathy) to macroalbuminuria (overt diabetic nephropathy) is age, where individuals diagnosed with diabetic nephropathy are older (54). This is in accordance with the mean of age of both groups in this study, where the case group presented a higher mean of age (67.8 years) compared with the control group (62.2 years). It is also well known that blood pressure increases with age, and systemic hypertension is a risk factor for the initiation and progression of diabetic nephropathy (54,55). Regarding the covariate diasease duration, its mean in this study was also higher in individuals from the case group (16.5 years) in comparison to controls (14.7 years). Therefore, a correlation between age and disease duration with the development of diabetic nephropathy can be found, since the case group is the one that presents individuals diagnosed with diabetic nephropathy, as well as older individuals and individuals with longer disease duration. The measurement of HbA1c levels is used to monitor the degree of control regarding glucose metabolism in diabetic individuals. In the normal 120-day lifespan of the red blood cells, glucose molecules react with hemoglobin through a posttranslational, non-enzymatic, substrate-concentration dependent irreversible process, in which the aldehyde group of glucose binds with the N-terminal of valine in the β-chain of hemoglobin, forming glycated hemoglobin. The International Diabetes Federation and American College of Endocrinology recommends a HbA1c value below 6.5%, while the American Diabetes Association recommends that the HbA1c be below 7.0% for most patients (157). In this study, the means for HbA1c in the control group (9.6%) as well as in the case group (8.7%) are above the recommended values by the International Diabetes Federation and American College of Endocrinology and the American Diabetes Association. This indicates that most of the individuals present in each of the groups have poor glycemic control. Clinical trials have demonstrated that a normalization of glycemia can greatly reduce the incidence of the diabetic complications associated with T2D.

However, in clinical practice, normalizing the blood glucose levels is not a trivial task, and almost 50% of diabetic subjects fail to reach the recommended HbA1c values (158). In respect to the covariate sex, it can be observed that the controls present more female sex individuals than males (52.6% against 47.4%) and the majority of individuals in the case group are male rather than female (64.7% against 35.3%).

Lastly and curiously, the study participant corresponding to exome 136, a 70 years old man, with a disease duration of 22 years and a HbA1c of 8.5% (Chapter 6| Appendices – Appendix B – Table B1), even though it possess all the determinants for having an increased risk of developing diabetic nephropathy, this individual does not present the diabetic complication, being part of the control group. However, the participant corresponding to exome 132, a 56 years old woman, with a disease duration of only 2 years and a HbA1c of 5.4% (Chapter 6| Appendices – Appendix B – Table B2) that do not present any of the factors responsible for an increased risk of developing diabetic nephropathy, presents the complication, being part of the case group. These observations support the idea of a genetic contribution for the development of diabetic nephropathy.

2. Whole-Exome Sequencing

2.1. DNA extraction and Quality Control

After the DNA extraction and before the initiation of the sequencing process, it is crucial to verify DNA integrity of all the samples that will be sequenced. That integrity was evaluated on an agarose gel. An example of the evaluation of sample DNA integrity is presented in Figure 22. The samples run in the gel were from exome 1 to exome 10, being exomes 2,3,4 and 7 the only ones that were included in this study.

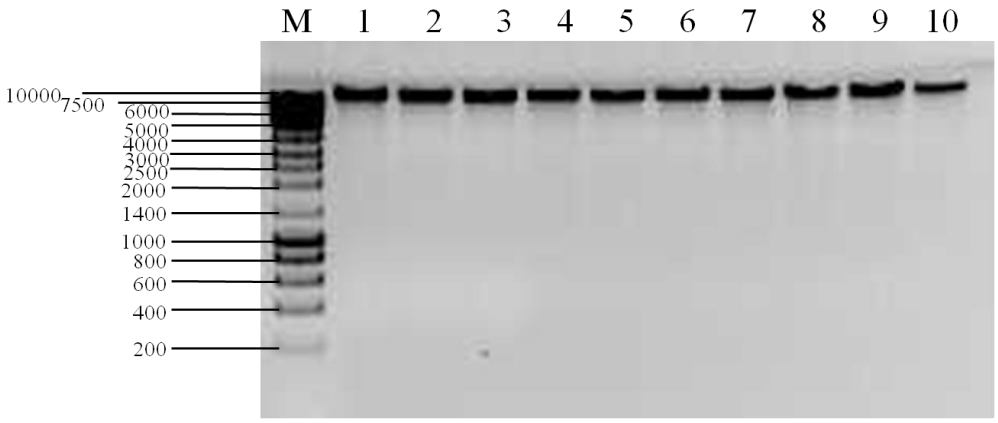

Figure 22. Electrophoresis of DNA samples in a 1% (w/v) agarose gel, using the molecular marker NZYDNA Ladder III. The DNA samples (exomes) to which each band in the gel corresponds are presented above the gel, and the band size, in base-pairs, of the ladder is represented on the left side of the agarose gel.

In the step following DNA extraction, namely the exome library preparation, high molecular weight genomic DNA is required. Therefore, if the DNA had good integrity, only a high molecular weight band was visible in the gel. Regarding all the DNA samples of individuals participating in this study, the gels performed to evaluate DNA integrity showed the presence of only a high molecular weight band and no other bands for smaller molecular weights, meaning that all the DNA samples revealed high integrity, allowing their exome library preparation and sequencing.

2.2. Sequencing process

After the exome library preparation, quality control was performed by evaluating the library profile, which allows to assess DNA fragments size and to verify if primer-dimers were present. An example of the expected library profile from Bioanalyzer of two different exomes from individuals participating in this study is presented in Figure 23.

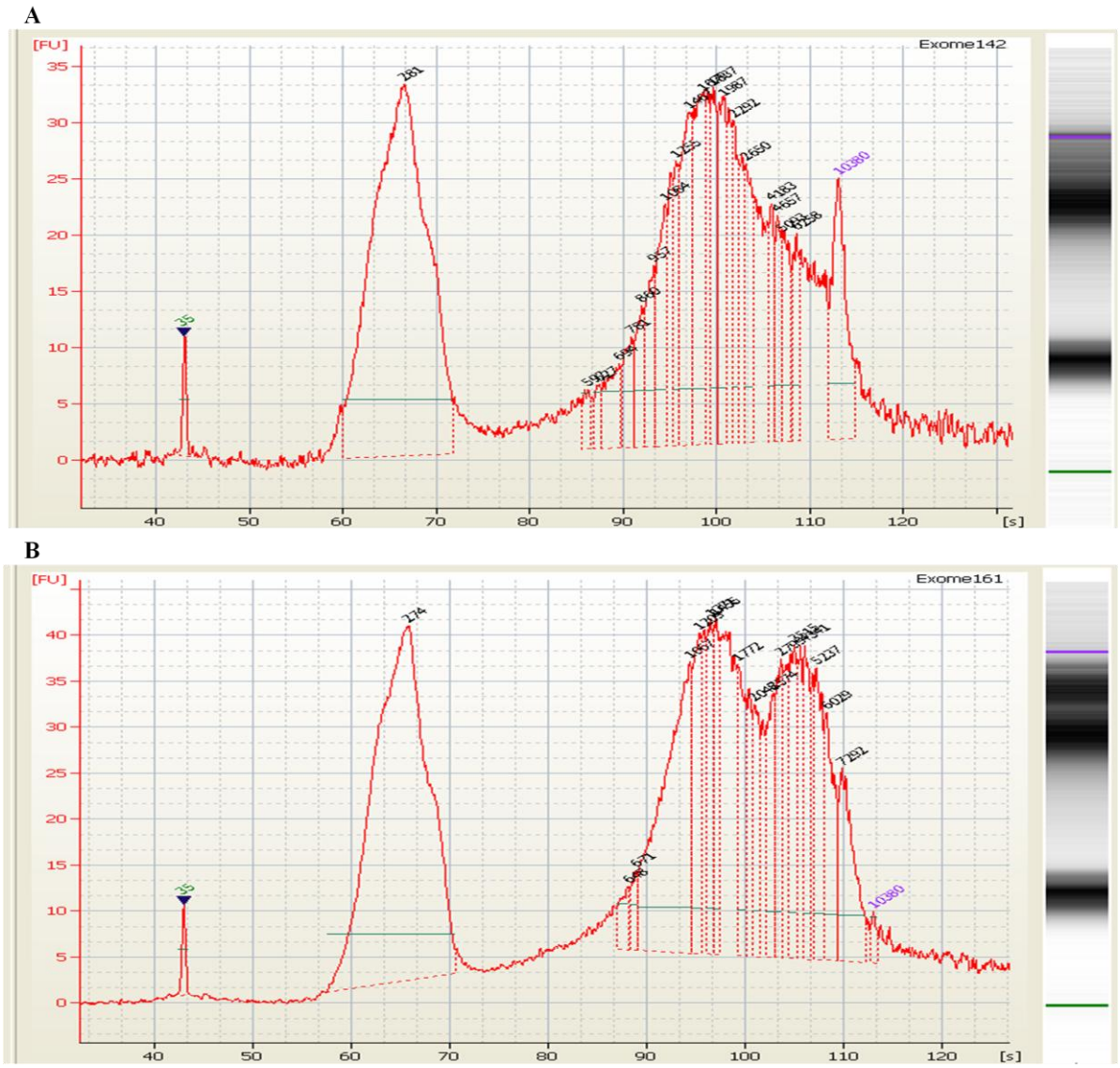

Figure 23. Expected exome library profile from Bioanalyzer. (A) library profile of exome 142, (B) profile of the library from exome 161. The black numbers on the peaks represent the size of the DNA fragments and the purple and green numbers represent the size of the markers (upper and lower markers, respectively).

The exome library quality is a critical determinant of the success of a sequencing run. The library profile obtained from Bioanalyzer for all the exome samples showed the peak corresponding to the expected size for the exon amplicons (around 200-300 bp), but also a larger peak with higher molecular weight that resulted from the overamplification during the PCR step (159). The larger amplicons are not included in the PCR emulsions and do not interfere with the sequencing. Primer-dimers (around 80-85 bp) can also be formed

during PCR, however, these PCR artifacts were not present in any of the 36 exome libraries. The library profile from all the individuals participating in this study was therefore the expected one, allowing the continuation of the sequencing process.

After each sequencing run in Ion Proton™ Sequencer several metrics are obtained for the sequenced exomes. In Figure 24, as an example, a sequencing run report of two exomes sequenced at the same time is displayed.

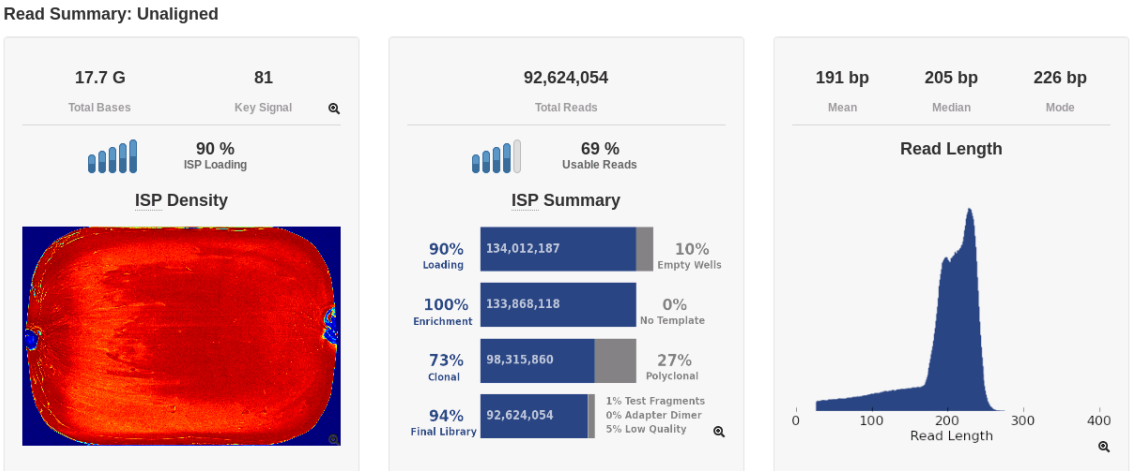

Figure 24. Sequencing run report from Ion ProtonTM Sequencer. This report refers to the run where exome 142 and exome 161 were sequenced.

A full report with the exome sequencing metrics is provided. The report example from exome 142 and exome 161 indicated a 90% chip load and a total of 92,624,052 reads with a mean read length of 191 bp, from both exomes combined. For the remaining 36 exomes, full reports for every 2 sequenced exomes were also obtained, with similar results to the ones presented here. The sequencing metrics obtained for the 36 exomes integrating this study were within the value range predefined by Life Technologies™ for exome sequencing. For a minimum coverage of 20x per base, the predefined values include a minimum of 30,000,000 reads per exome library and a minimum mean read length of 140 bp.

2.3. Bioinformatics Analysis

In WES, there is still a high variation in coverage, regardless of the sequencing platform used. Therefore, an average depth of coverage of 80x is required to cover 89.6% to 96.8% of the target region (160). The coverage analysis for each exome is presented in Appendix C – Table C1 and, in Table 12, only the mean and standard deviation of the coverage analysis from the reads obtained in the sequencing process is showed.

Table 12. Mean and standard deviation of the coverage analysis metrics.

Coverage Analysis metrics	Mean ± Standard Deviation
Mapped Reads (number)	40,674,699 ± 7,782,233
Reads on Target (%)	94.41 ± 0.01
Mean Depth (x)	116.24 ± 24.38
Uniformity (%)	92.42 ± 0.01

Many factors influence the minimum read depth of coverage that is required to address adequately a biological question using exome sequencing. A higher depth of coverage is usually associated with a more reliable calling of the genetic variants present in a given population, allowing a reduction of the false-discovery rate in the variant calling step (160). Therefore, the coverage analysis metrics obtained for the 36 exomes was, for the mapped reads, namely the number of reads aligned with the reference genome, a mean of 40,674,699 reads; the mean of the reads on target, reads that were aligned to the target region, was 94.41%; the mean of the mean depth of coverage, in other words, the average number of times that a particular nucleotide is represented in a collection of random sequences, was 116.24x and lastly, the coverage uniformity, meaning the "horizontal" coverage of the targeted genomic intervals, was 92.42%. Since a depth of coverage around 80x is the required for the coverage of 89.6% to 96.8% of the target region, it can be considered that all exomes in this study, with the exception of exome 3, 4 and 116, presented a good coverage of the intended target region (Chapter 6| Appendices – Appendix C – Table C1). The obtained coverage analysis metrics for the total of exomes studied were within the value range predefined by Life Technologies™ for exome sequencing. For a minimum coverage of 20x per base, the predefined values are a

minimum of 30,000,000 mapped reads per exome library and a minimum of 85% of reads on target and uniformity.

The mapping step in the bioinformatics analysis consisted in the alignment of the reads to the human genome reference sequence hg19. This ultimately allowed the execution of the variant calling step. In this step, all the positions in which the aligned sequence deviates from the reference sequence, creating genetic variants, are identified. All types of genetic variants present in the 36 exomes under study, as well as the number of homozygous and heterozygous for each variant type, are represented in Figure 25. The table from which the graphic representation originated, with the total number of genetic variants present in the study population, as well as the number of homozygous and heterozygous for each variant type by exome, is presented in Appendix C – Table C2.

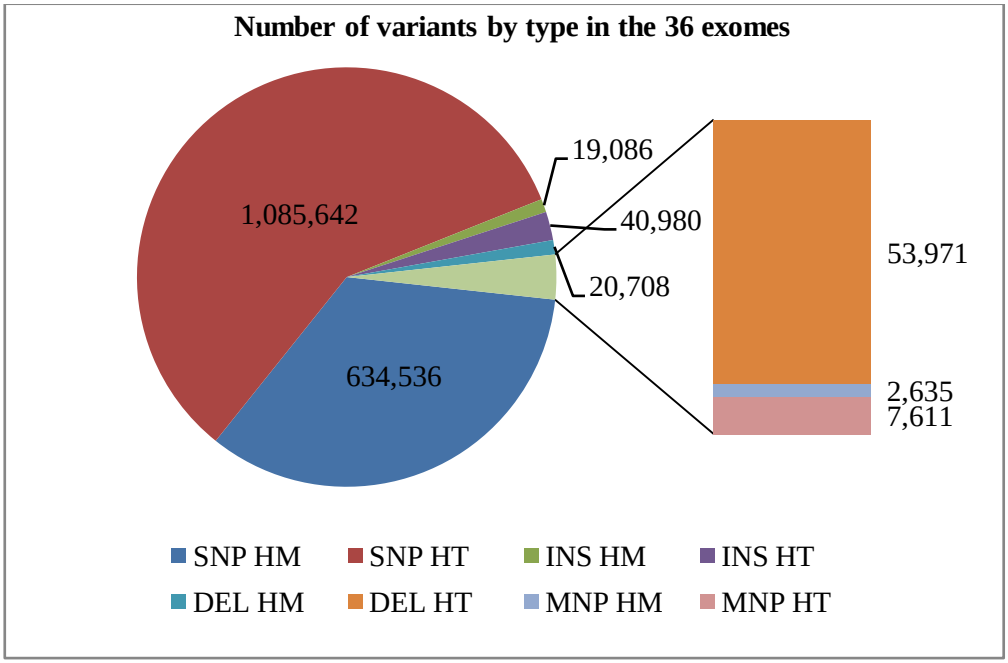

Figure 25. Graphic representation of the number of homozygous and heterozygous for each variant type in the 36 exomes. DEL HM: homozygous deletion; DEL HT: heterozygous deletion; INS HM: homozygous insertion; INS HT: heterozygous insertion; MNP HM: homozygous multiple nucleotide polymorphism; MNP HT: heterozygous multiple nucleotide polymorphism; SNP HM: homozygous single nucleotide polymorphism; SNP HT: heterozygous single nucleotide polymorphism.

In the 36 exomes there was a total of 634,536 homozygous and 1,085,642 heterozygous SNPs, 19,086 homozygous and 40,980 heterozygous insertions, 20,708 homozygous and 53,971 heterozygous deletions and lastly a total of 2,635 homozygous and 7,611 heterozygous multiple nucleotide polymorphisms (MNPs). In general, there were more heterozygous genotypes for each variant type than homozygous. This latter genotype represented a combination of wild-types and altered homozygous. Furthermore, MNPs were, by far, the less common variant type found in the study population, while SNPs were the main type of variant found, being also the type of variant in which this work focused.

3. Common Variants Approach

3.1. Statistical Analysis

As previously stated, it has been proposed that complex diseases be, in part, determined by genetic factors. This suggestion made it essential to identify the genetic basis for all complex diseases, and several methods emerged to study those bases (9,19). With an ever increasing number of methods developing, it became possible to identify and associate common variants with complex diseases, being a large number of those variants, primarily SNPs, already identified and associated with most of the complex human diseases and related traits (104).

In this study, a list of common variants was obtained from the statistical analysis. Since the number of statistically significant common variants obtained, those with a p-value $\leq$ 0.05, was high (654 common variants), several filters were applied to prioritize the variants. Therefore, the variants present in the 3' or 5'-UTR, those intronic, the ones upstream or downstream of the gene, as well as the synonymous variants, had to have a CADD (135) > 15 and be in conserved regions. Missense variants had to have at least one of the prediction programs (PolyPhen-2 (132), SIFT (134), and CADD) indicating that the variant was pathogenic or the variant had to be in a conserved region. In Appendix D – Table D1, a list of all the filtered genetic variants (135 common variants) is presented.

Those filtered common variants were then studied based on the literature, to detect the most biologically relevant to the pathogenesis of diabetic nephropathy. Therefore, with this approach, 6 common variants present in 5 different genes were found as the most relevant

to the pathogenesis of diabetic nephropathy. Of those, 4 variants (rs1051303 and rs1131620 in *LTBP4*, rs660339 in *UCP2* and rs2589156 in *RPTOR*) were considered protective variants, presenting an odds ratio (OR) < 1 and being primarily present in the control group. The remaining 2 variants (rs2304483 in *SLC12A3* and rs10169718 in *ARPC2*) were considered risk variants, presenting an OR > 1 and being mainly present in the case group. An increased frequency of an altered allele or genotype in the case group compared to the control group indicates that the altered allele or genotype may be involved in an increased risk for developing the disease being studied (156).

The annotation of the relevant common variants is presented in Table 13. The software applications PolyPhen-2, SIFT, and CADD indicated the impact of the SNP variants in the coding regions, while the programs SplicePort: An Interactive Splice Site Analysis Tool (161), Human Splicing Finder (HSF) (162), Analyzer Splice Tool, Splice Site Prediction by Neural Network (NNSplice) (163), HBond Score Web-Interface (164) and USD SplicePredictor Online Service (165) classified genetic variants in the splice regions. The results from those programs are available in Appendix D – Table D2.

Table 13. Annotation of the common variants biologically relevant to diabetic nephropathy.

	Gene	rs ID	Ref. allele	Alt. allele	MAF (cases)	MAF (controls)	MAF (EUR) ESP/ 1000G	Type of variant	PolyPhen-2 *	SIFT **	CADD ***	p-value	Odds ratio (95% CI)	Associated mechanism
Protective	*LTBP4*	rs1051303	A	G	0.29	0.68	0.42/0.45	Missense	Benign	Tolerated	0.07	0.03	0.21 (0.051-0.853)	Regulation of TGF-β release
		rs1131620	A	G	0.31	0.68	0.42/0.45	Missense	Benign	Tolerated	0.00	0.05	0.30 (0.091-0.997)	
	UCP2	rs660339	G	A	0.26	0.58	0.41/0.42	Missense	Benign	Tolerated	9.07	0.02	0.07 (0.006-0.656)	Oxidative stress
	RPTOR	rs2589156	G	A	0.03	0.26	0.13/0.10	Splice	-	-	10.18	0.01	0.04 (0.003-0.460)	mTOR signaling pathway
Risk	*SLC12A3*	rs2304483	T	C	0.68	0.42	0.38/0.42	Splice	-	-	2.18	0.01	18.40 (1.973-171.566)	Systemic hypertension and RAAS pathway
	ARPC2	rs10169718	A	G	0.62	0.32	0.49/0.53	Splice	-	-	6.68	0.03	7.83 (1.284-47.735)	Actin polymerization

1000G: 1000 genomes project; Alt.: altered; CADD: combined annotation dependent depletion; CI: confidence interval; ESP: exome sequencing project; EUR: Europe; MAF: minor allele frequency; mTOR: mammalian target of rapamycin; PolyPhen-2: polymorphism phenotyping v2; RAAS: renin-angiotensin-aldosterone system; Ref.: reference; SIFT: sorting intolerant from tolerant; TGF-β: transforming growth factor β.

*PolyPhen-2 (132): "benign" (≥ 0 and ≤ 0.452); "possibly damaging" (≥ 0.453 and ≤ 0.956) and "probably damaging" (≥ 0.957 and ≤ 1)

**SIFT (134): "deleterious" (≤ 0.05) and "tolerated" (> 0.05)

***CADD (135): higher scores corresponds to a higher pathogenicity (in this study a genetic variant was considered pathogenic with a score ≥ 12)

The variants rs2589156 in the *RPTOR* gene, rs2304483 in *SLC12A3* and rs10169718 in the *ARPC2*, are splice region variants. For the variant present in *RPTOR*, there is a possible loss of the splice region, since the scores obtained from the prediction programs for the sequence with the variant were well below the ones obtained for the reference sequence. This variant is present in a donor splice site, which implies a possibility for the intron to start being considered an exon. For the remaining variants, rs2304483 in *SLC12A3* and rs10169718 in the *ARPC2*, not all the prediction programs are in accordance. However, the majority considered these variants benign ones, since the scores obtained for the sequence with the variant are similar to the ones obtained for the reference sequence. This suggests that the variants do not affect the splice region. Furthermore, all of these common variants presented the minor alleles as altered alleles, with the exception of the variant rs2304483 in the *SLC12A3* gene. In this variant, the minor allele corresponds to the reference allele.

The genetic variants found in this approach are associated with several different mechanisms. The variants rs1051303 and rs1131620 present in *LTBP4* gene play a role in releasing and activating TGF-β, the variant rs660339 *UCP2* contributes for oxidative stress, the rs2589156 in *RPTOR* is involved in the mammalian target of rapamycin (mTOR) signaling pathway, the genetic variant rs2304483 in the gene *SLC12A3* can be related to systemic hypertension and appears to be involved in RAAS pathway, and lastly, the variant rs10169718 present in the gene *ARPC2* is implicated in actin polymerization.

There is accumulating evidence that suggest that the TGF-β system plays an important role in the pathogenesis of diabetic nephropathy by regulating ECM proteins metabolism (mainly regulated by the TGF-β1 isoform). This factor is secreted in a large complex with no biological activity that consists of TGF-β, LAP and LTBP, being cleavage of both LAP and LTBP necessary for TGF-β activation (166). Like the others LTBPs, the LTBP4, encoded by the *LTBP4* gene, is the constituent of the large latent complex responsible for TGF-β1 transportation into the ECM, limiting its availability to interact with TGF-β receptors (type II and type I serine/threonine kinase receptors). Thereby, proteolytic cleavage of LTBP4 allows the release of TGF-β1 from the ECM, rendering it accessible to connect with TGF-β receptors and consequently exert its effects on the ECM proteins metabolism through the Smad proteins (Figure 26) (167).

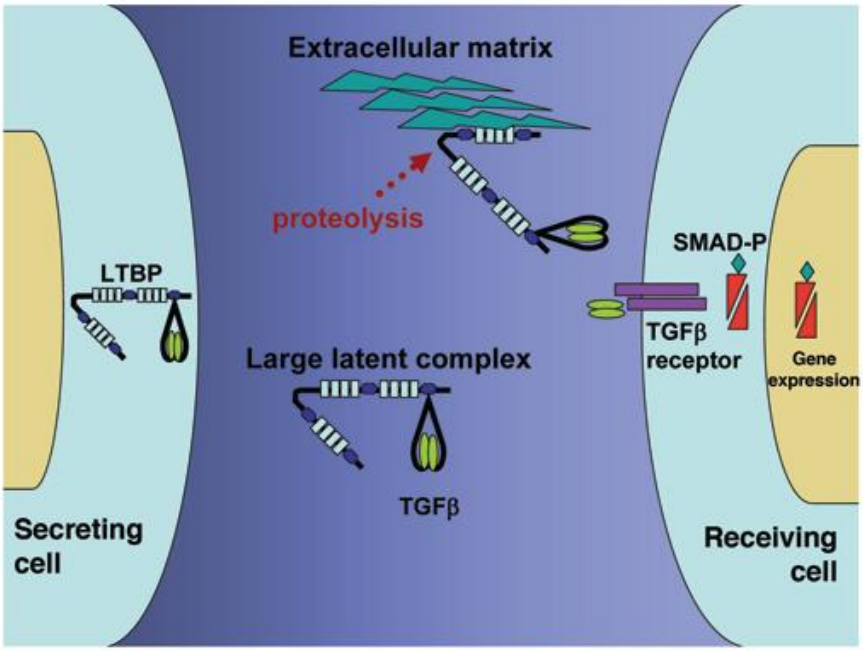

Figure 26. Model for LTBP4 action (taken from (167)). LTBP: latent TGF-β binding protein; SMAD-P: Smad proteins phosphorylation; TGF-β: transforming growth factor β.

The LTBPs are produced in excess to TGF-β, and since TGF-β secretion is very inefficient in the absence of LTBPs, it is probable that most secreted cellular TGF-β is in large latent complexes (78). Therefore, an enhanced proteolytic susceptibility of LTBP4 is associated with increased TGF-β1 release and consequent Smad activation, being LTBP4 responsible for the regulation of TGF-β secretion and activation (166,167). Furthermore, both of these variants, as well as others present in the *LTBP4* gene, were also associated to Duchenne muscular dystrophy (DMD), the most common muscular dystrophy characterized by progressive muscle degeneration. This disease course varies among patients, but most of those patients lose ambulation between 6 and 12 years of age. The reason behind the differences in disease progression is still not clear. However, genetic modifiers in the *LTBP4* gene have been proposed to contribute to those differences. This study shows a significant association between four SNPs in the *LTBP4* gene, including the two variants associated to diabetic nephropathy in this study, and DMD progression. The four SNPs were associated with prolonged ambulation due to reduced TGF-β signalling. TGF-β is known to increase fibrosis in muscle tissue and several studies have shown that inhibition of its function decreases the presence of fibrosis and increases muscle strength in mdx mice, a surrogate for DMD (168).

The gene *UCP2*, where the common variant rs660339 is present, is associated with ROS production. Nowadays, there is increasing evidence that the overproduction of ROS is one of the major factors responsible for the development of diabetes itself as well as diabetic complications (169). Excessive amounts of ROS, mainly from the mitochondrial electron transport chain and nicotinamide adenine dinucleotide (NADH)/NADPH oxidases, after surpassing various endogenous anti-oxidative defensive mechanisms, oxidize various tissue biomolecules, such as, DNA, proteins, carbohydrates and lipids, being this state commonly referred to as oxidative stress (42). In diabetic nephropathy, ROS, in addition to mediate high glucose-induced angiotensinogen activation consequently increasing angiotensin II and contributing for RAAS over-activity, also induce TGF-β1 overexpression in mesangial cells. It is known that a combination of strategies to prevent the overproduction of ROS (good glycemic control and/or inhibition of cytokines and growth factors) and to increase the removal of pre-formed ROS (conventional or catalytic antioxidants) may prove to be effective in preventing the development and progression of diabetic nephropathy (169).

The uncoupling proteins (UCPs) belong to the mitochondrial anion transporter superfamily located in the inner mitochondrial membrane, having been described UCP homologs from 1-5 in mammals. It is well known that the protein UCP2, encoded by the *UCP2* gene, is a mitochondrial antioxidant protein, whose inhibition induces oxidative stress favoring the formation of mitochondrial superoxide (170). Superoxide is the initial oxygen free radical formed by the mitochondria, which is then converted to other more reactive species that can damage cells in numerous ways. Normally, electron transfer through complexes I, III, and IV extrudes protons outward into the intermembrane space, generating a proton gradient that drives ATP synthase (complex V) as protons pass back through the inner membrane into the matrix. In diabetic cells, due to the high intracellular glucose concentration, there is more glucose-derived pyruvate, which generates more electrons donors (NADH and flavin adenine dinucleotide ($FADH_2$)), resulting in an increased flux of electrons in the electron transport chain and consequently increased proton gradient across the inner mitochondrial membrane. As a result of this increase, electron transfer inside complex III is blocked, causing the electrons to back up to coenzyme Q, which donates the electrons one at a time to molecular oxygen, thereby generating superoxide. The mitochondrial isoform of the enzyme superoxide dismutase

(SOD) degrades this oxygen free radical to hydrogen peroxide, which is then converted to water (H_2O) and oxygen (O_2) by other enzymes (Figure 27) (171).

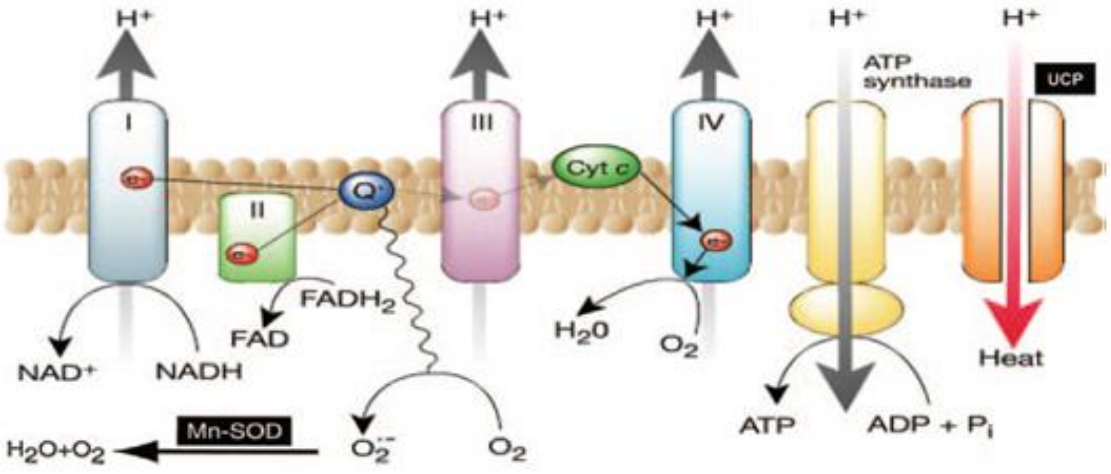

Figure 27. Production of ROS by the mitochondrial electron transport chain (adapted from (171)). ADP: adenosine diphosphate; ATP: adenosine triphosphate; Cyt c: cytochrome c; FAD/$FADH_2$: flavin adenine dinucleotide; H^+: hydrogen ion; H_2O: water; Mn-SOD: manganese-dependent superoxide dismutase; NAD^+/NADH: nicotinamide adenine dinucleotide; O_2: oxygen; $O_2^{\cdot -}$: superoxide; Pi: inorganic phosphate; Q: coenzyme Q; UCP: uncoupling protein.

As mentioned above, the passage of electrons through the chain generates a proton gradient across the membrane, being these protons allowed passage either via ATP synthase, which drives oxidative phosphorylation and consequent ATP production, or via UCPs, which dissipate the energy of the proton gradient as heat (44,171). It is then clear that UCPs increase proton conductance of the mitochondrial inner membrane, but only when they are activated by products of ROS metabolism, such as hydroxynonenal. In their absence, they do not affect the basal proton conductance of the membrane. Therefore, it is thought that superoxide production from the electron transport chain contributes to UCPs activation, being in this particular case, the activation of protein UCP2 (Figure 28) (172).

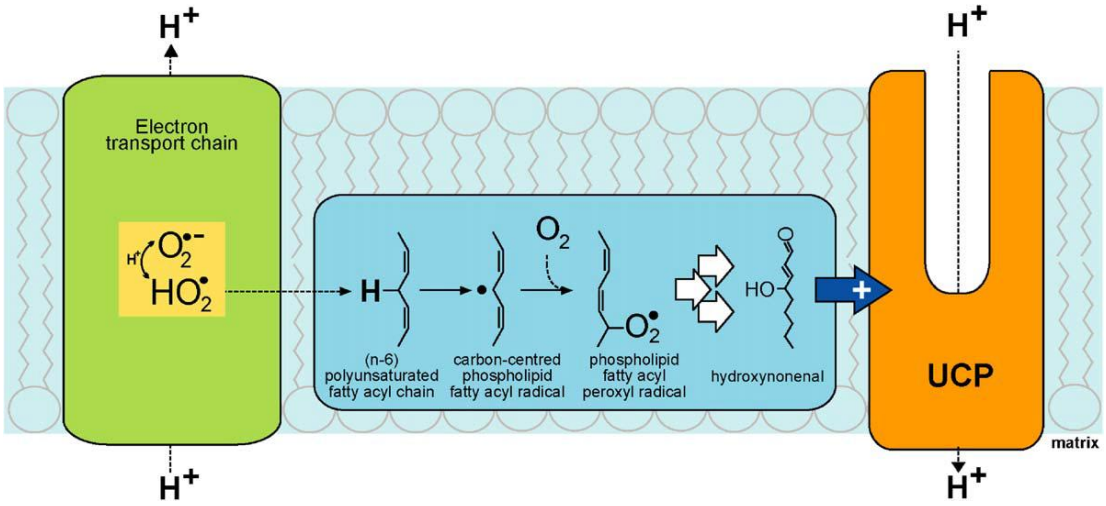

Figure 28. Model for the activation of the protein UCP2 (taken from (172)). H: hydrogen atom; H^+: hydrogen ion; $HO_2^{\cdot}$: hydroperoxyl radical; O_2: oxygen; $O_2^{\cdot -}$: superoxide; OH: hydroxide; UCP: uncoupling protein.

Superoxide, or its more lipid-soluble protonated form, hydroperoxyl radical, diffuses into the inner membrane, where it attacks *n*-6 polyunsaturated fatty acyl chains of membrane phospholipids, such as arachidonate, extracting a hydrogen atom and leaving behind a carbon-centered fatty acyl radical. This carbon-centered fatty acyl radical can react with molecular oxygen to form a lipid peroxyl radical, which initiates another round of carbon-centered radical production and breaks down to fragments like 4-hydroxynonenal. In this way, a single hydroperoxyl radical can initiate a cascade that generates a large number of hydroxynonenal molecules that will activate, in this particular case, UCP2 (172). Activated UCP2 mildly uncouples substrate oxidation from ATP synthesis, thereby dissipating a portion of the mitochondrial proton gradient and, consequently, decreasing ATP production by the electron transport chain (173). Therefore, this uncoupling renders the electron flow through the electron transport chain complexes more efficient and decreases mitochondrial superoxide production, ultimately leading to a decreased ROS formation by mitochondria (170,173). The attenuation of mitochondrial ROS production eventually protects against ROS-related cellular damage (172).

Podocytes play a critical role in preventing plasma proteins from leaking into the urine, being essential in the glomerular filtration barrier. Therefore, any kind of podocyte injury is believed to contribute to the development of diabetic nephropathy (174). The mTOR is an evolutionarily conserved protein kinase that plays a major role in cell growth and cell

size control. This serine/threonine kinase is the catalytical subunit of two multi-protein complexes, the mTOR complex 1 (mTORC1) and the mTOR complex 2 (mTORC2) (174,175,176). These complexes can be distinguished by their unique composition and different substrates. The mTORC1 is composed of five elements, the mTOR, its catalytic subunit, the regulatory-associated protein of mTOR, complex 1 (RPTOR, also known as Raptor), the mammalian lethal with Sec13 protein 8 (mLST8, also known as GβL), the proline-rich AKT substrate 40 kDa (PRAS40) and the DEP-domain-containing mTOR-interacting protein (Deptor), being their exact function still elusive. The mTORC2 comprises six different proteins, several of which are common to mTORC1 and mTORC2 (176,177). The mTORC1 is rapamycin-sensitive and regulates a wide array of cellular processes including cell growth, proliferation, and autophagy, in response to nutrients such as insulin, glucose and amino acids. In response to these stimuli, mTORC1 is activated by two families of Ras-related small guanosine triphosphate (GTP)ases, the Ras homolog enriched in brain (Rheb) and Rags, even though the precise molecular mechanisms by which they activate mTORC1 remain unclear (175,178). Growing evidence suggests that increased activation of mTORC1 is likely to be associated with the pathogenesis of diabetic complications, being mTORC1 pathologically activated in podocytes in diabetic nephropathy early stages and its inactivation responsible for ameliorating podocyte injury. Furthermore, it was also found that high mTORC1 activation led to mislocalization of nephrin, a podocyte cell surface protein, causing it to be retained in the cytoplasm and, therefore, preventing the correct formation of the filtration barrier (Figure 29) (174,175).

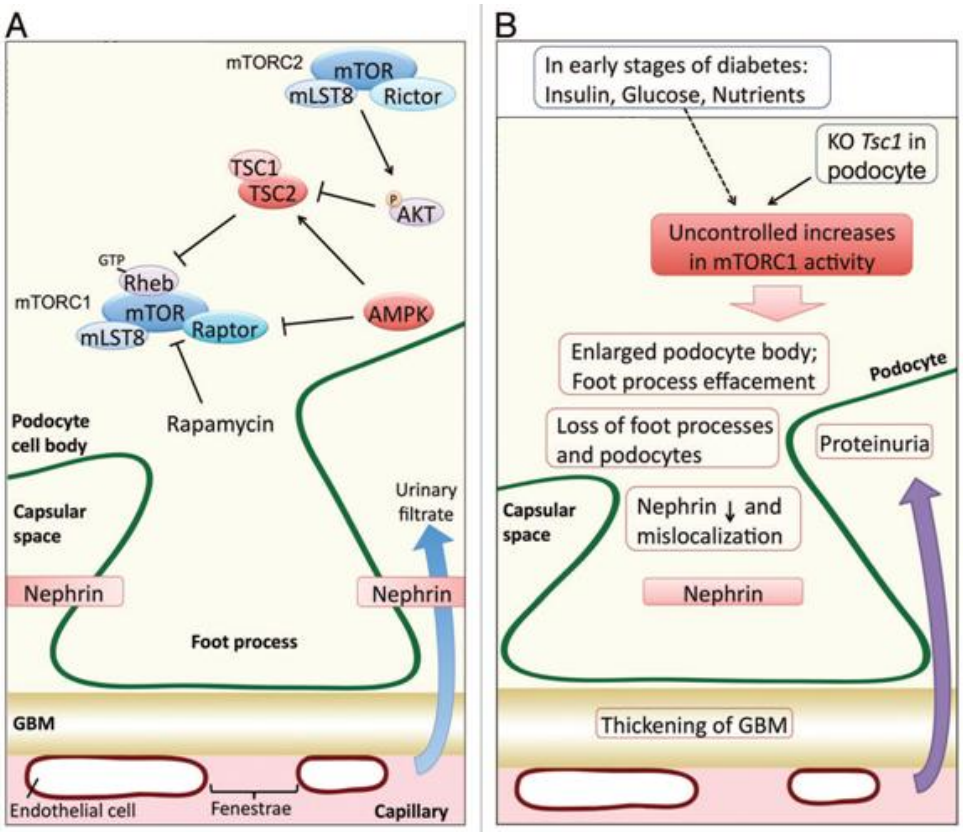

Figure 29. mTOR signaling and nephrin localization in podocytes (adapted from (174)). In (A) is shown the mTOR signaling and nephrin localization in normal podocytes, and in (B) is shown the pathological mTORC1 activation that contributes to podocyte injury. AKT: alpha serine/threonine protein kinase; AMPK: adenosine 3′,5′-monophosphate-activated protein kinase; GBM: glomerular basement membrane; GTP: guanosine-5'-triphosphate; mLST8: mammalian lethal with Sec13 protein 8; mTOR: mammalian target of rapamycin; mTORC1: mammalian target of rapamycin complex 1; mTORC2: mammalian target of rapamycin complex 2; Raptor: also known as RPTOR (regulatory-associated protein of mTOR, complex 1); Rheb: Ras homolog enriched in brain; Rictor: RPTOR independent companion of mTOR, complex 2; TSC1: tuberous sclerosis complex 1; TSC2: tuberous sclerosis complex 2.

Increased activation of mTORC1 in podocytes resulted in proteinuria, loss of podocytes and foot processes, foot processes effacement and changes in the components of the glomerular basement membrane. Therefore, this complex inhibition has recently been suggested as a new therapeutic target to prevent diabetic nephropathy. However, the inhibition of mTORC1 also causes podocyte dysfunction leading to proteinuria, which suggests that a basal level of mTORC1 activity is required for maintaining the basic

physiology of podocytes. So, it is speculated that, under normal conditions, mTOR activity is essential for normal podocyte development and function (45,174,175). Although the precise molecular mechanisms of how mTORC1 deregulation might affect podocyte needs to be delineated, it is thought that the increase in its activity associated with diabetic nephropathy is likely an attempt to maintain podocyte homeostasis. However, this activation, which may provide some short-term benefits, ultimately causes this complex to mediate adverse effects (45,176). The protein encoded by the *RPTOR* gene, the gene in which the common variant found in this study population as associated with diabetic nephropathy, is an essential component of mTORC1, affecting this complex activity by regulating its assembly and recruiting the necessary substrates. Furthermore, this protein also plays a crucial role in determining the subcellular localization of mTORC1 and is the target of rapamycin, the mTORC1 inhibitor that performs its function by enhancing the dissociation of this protein from the complex, ultimately resulting in mTORC1 inhibition, which, as mentioned above causes podocyte dysfunction leading to proteinuria (174,179,180).

The variant rs2304483 in the *SLC12A3* gene has not yet been reported as associated with diabetic nephropathy. However, another common variant, rs11643718, also present in this gene, was associated in a Malaysian (181) as well as in a Japanese population (182) of type 2 diabetic individuals, as having a protective effect in diabetic nephropathy, therefore reducing the risk to develop this complication. Furthermore, another study in a type 2 diabetic Korean population implicated an association between this same genetic variant and an increased risk for developing ESRD caused by diabetic nephropathy (183). On the other hand, there was also a study in which no association was found between this variant and diabetic nephropathy among a type 2 diabetic American Caucasians population (184). These studies performed so far present conflicting data, being one hypothesis for explaining that data divergence the fact that the effect of *SLC12A3* genetic variants could be population specific (181,184). Therefore, it is necessary to carry further studies to clarify the effects of genetic variants present in the *SLC12A3* gene in diabetic nephropathy (181).

So far, little is known about the biological function of this gene, being only recognized that *SLC12A3* encodes a thiazide-sensitive sodium/chloride (Na^+/Cl^-) co-transporter in the kidney. This co-transporter is responsible for the reabsorption of sodium (Na^+) and

chloride (Cl^-) from the lumen into the kidney endothelial cell in the distal convoluted tubule, being Na^+ then returned to the blood system via the ATP-dependent sodium-potassium pump (181,182). This transporter is thiazine-sensitive, being a hypothesis for the connection between this gene biological function and diabetic nephropathy that, as thiazide is widely used for systemic hypertension treatment, a known risk factor for the development and progression of diabetic nephropathy, a loss of function of this co-transporter could be associated with a reduced blood pressure. Alternatively, this transporter could also regulate the reabsorption of unknown molecules, being a sustained decrease in its activity responsible for the accumulation of nephrotoxic substances (182). Furthermore, loss of function mutations in *SLC12A3* are responsible for Gitelman syndrome, an inherited autosomal recessive trait characterized by low blood pressure due to renal Na^+ wasting and electrolyte abnormalities, such as hypokalemia, metabolic alkalosis and hypocalciuria (182,183). This gene is also involved in the RAAS pathway in the lumen of a kidney cell. However, its precise role in this pathway still needs to be explored (181).

The biological function of the *ARPC2* gene is associated with a component of the glomerular filtration barrier, the podocytes. Over the years, it has become clear that podocytes are structurally and functionally complex cells that play a key role in preventing proteinuria, being a correct organization and regulation of the actin cytoskeleton in the podocyte essential for the maintenance of its morphology and consequently their function (185). In diabetic nephropathy, there can be a disruption of the actin cytoskeleton due to the impairment or collapse of actin filaments, which may cause podocyte foot process effacement and the emergence of proteinuria (185,186).

It is now recognized that the nephrin-neph1-podocin receptor complex interact with actin cytoskeleton associated proteins, thereby signaling to regulate foot process cytoskeletal dynamics and morphology (92). This receptor complex is capable of assembling the protein complex responsible for inducing polymerization and elongation of the actin filaments, the actin-related protein-2/3 (ARP2/3) complex. Once the receptor complex is activated by tyrosine phosphorylation, it recruits the non-catalytic region of tyrosine kinase adaptor protein (Nck) and the growth factor receptor-bound protein 2 (Grb2) to initiate actin polymerization. The first protein serves as an adaptor protein that recruits actin associated proteins such as neural Wiskott-Aldrich syndrome protein (N-

WASP), components of the ARP2/3 complex and other components of the actin polymerization machinery that are necessary for its induction and regulation (187). The actin filaments are constituted by a fast growing end (barbed end) and a slower growing end (pointed end) and the dynamic assembly and disassembly of those filaments are crucial aspects of the actin function, being controlled by over a hundred actin-binding proteins. These proteins bind directly to filaments or monomers and control actin structure and dynamics by nucleating, capping, stabilizing, severing, depolymerizing, crosslinking, bundling, sequestering or delivering monomers, or by promoting monomer nucleotide exchange. An important set of actin regulators initiates formation of new actin filaments by a process called nucleation.

The spontaneous nucleation is a kinetic hurdle in the process of actin polymerization, and, therefore, factors that can accelerate or bypass this step are important for efficient actin assembly in the cell. So far, three classes of protein have been identified that initiate new filament polymerization: the ARP2/3 complex, the formins, and spire (188). The ARP2/3 complex was the first of these molecules to be identified and has since been shown to have a crucial role in the formation of branched-actin-filament networks (186,188). This complex consists of a stable assembly of seven polypeptides, the actin-related protein 2 and 3 (ARP2 and ARP3) and the remaining subunits are actin-related protein complex 1 to 5 (ARPC1, ARPC2, ARPC3, ARPC4 and ARPC5). The ARP2/3 complex possesses little biochemical activity on its own. However, when engaged by nucleation-promoting factor (NPF) proteins, it is activated to initiate the formation of a new (daughter) filament that emerges from an existing (mother) filament in a y-branch configuration with a regular 70° branch angle (Figure 30). The actin polymerization by this complex is autocatalytic, being the manner by which the nucleation and branching activities are linked together still a matter of debate (188).

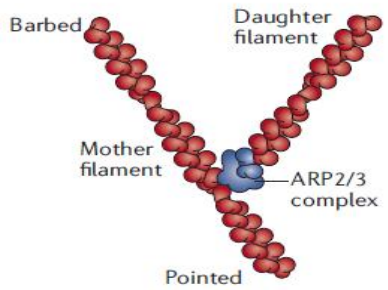

Figure 30. Diagram of ARP2/3 complex (adapted from (188)). ARP2/3 complex: actin-related protein-2/3 complex.

The protein ARPC2, encoded by the gene *ARPC2*, along with ARPC4 form the structural core of the complex, being the remaining subunits organized around them. The proposed models for this complex structure share the common principles that ARP2 and ARP3 interact with the pointed end of the daughter filament while the proteins ARPC2 and ARPC4 make substantial contacts with the mother filament (188).

Furthermore, the biologically relevant common variants were localized in their respective genes to determine if some gene presented a variant "hotspot" (Figure 31).

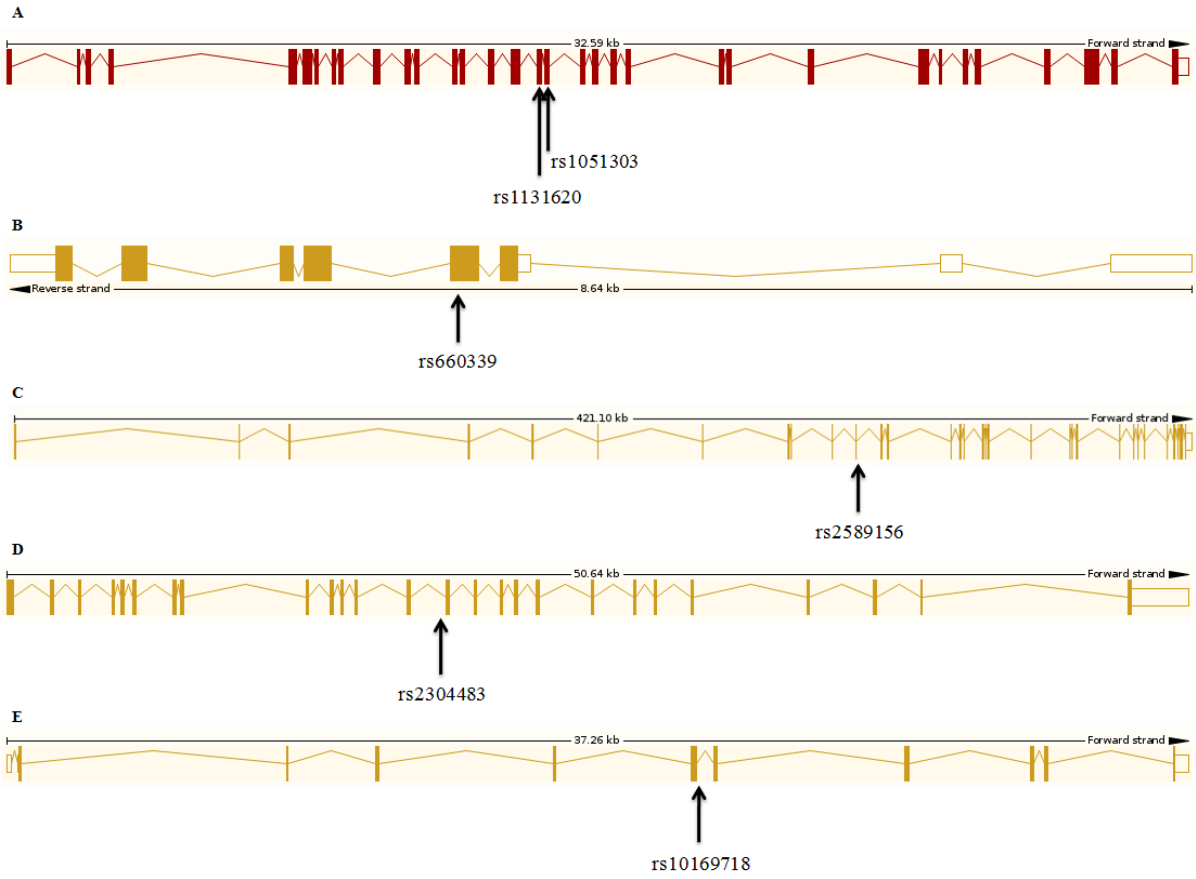

Figure 31. Localization of the common variants in their respective genes. (A) two variants present in *LTBP4*, (B) variant present in *UCP2*, (C) variant in *RPTOR*, (D) variant present in *SLC12A3* and (E) variant localized in *ARPC2*. The gene sequences used were from version GRCh37 of Ensembl (139).

The presence of a variant "hotspot" could not be determined in any of the genes, since the number of genetic variants present in each gene was reduced. However, in the *LTBP4* gene, two statistically significant common variants were found. Their localization, even though not enough to determine if there is or not a "hotspot" in this gene, can provide a

general idea of the subject. Therefore, those variants are located in neighboring exons of the gene, allowing the speculation of whether not an exon but the region surrounding them can be considered a variant "hotspot". Nonetheless, in order to determine a "hotspot" in any gene with a certain degree of certainty, the presence of a significant number of genetic variants in that same gene is required.

The missense genetic variants also have an impact regarding proteins, since they are localized in the genes coding regions (exons). The changes in amino acids that occur as a consequence of those genetic variants are presented in Table 14.

Table 14. Functional impact of the missense common genetic variants.

Gene	rs ID	Ref. amino acid	Alt. amino acid	Change position/Protein length	Codon change
LTBP4	rs1051303	Threonine	Alanine	273/899	Acc/Gcc
	rs1131620	Threonine	Alanine	240/899	Act/Gct
UCP2	rs660339	Alanine	Valine	55/225	gCc/gTc

Alt.: altered; Ref.: reference.

The structure for the LTBPs proteins is highly repetitive. They are mainly composed of epidermal growth factor (EGF)–like repeats, eight cysteine (8-Cys) repeats, as well as flanking regions containing proline-rich areas, or in some cases, glycine-rich areas. The N-terminal of all LTBPs contains two or three copies of EGF-like repeats, one 8-Cys repeat, and another 8-Cys repeat often called the hybrid domain. That hybrid domain contains seven cysteines and shares similarity with both EGF and 8-Cys repeats (78). The N-terminal of these proteins is responsible for the interactions between the LTBPs and the ECM. In the case of the LTBP4 protein, following the N-terminal domain, there is a protein sensitive region, called the hinge region, which is rich in proline and basic amino acid residues. Proteolytic cleavage at this region is thought to be responsible for the latent TGF-β release from the ECM and activation (78,167). The central part of all the LTBPs is a region resistant to proteolysis that consists of 9-14 EGF-like repeats and is supposed to form a helical rod-like structure. Moreover, these EGF-like repeats participate in protein-protein interactions in addition to providing stability to protein structures via calcium binding (78). At the C-terminal of all LTBPs and consequently also in LTBP4, the 8-Cys

repeats bind the small latent TGF-β complex, forming the large latent complex (78,167). The LTBP4 protein structure is shown in Figure 32.

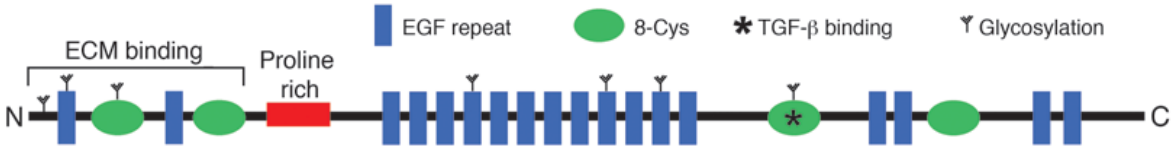

Figure 32. Protein structure for LTBP4 (adapted from (167)). 8-Cys: eight cysteine; ECM: extracellular matrix; EGF: epidermal growth factor; TGF-β: transforming growth factor β.

Both variants present in the *LTBP4* gene, rs1051303 and rs1131620, cause a change from the amino acid threonine to alanine at the protein positions 273 and 240, respectively. This change switches an essential amino acid characterized by a polar side chain (threonine) for a non-essential amino acid with a nonpolar side chain (alanine) in an interspecies conserved region for both variants. This change, theoretically, leads to a higher functional impact. Based on the information provided by UniProtKB, it can be verified that both of the amino acid changes occur between domains, not directly affecting neither of them. Therefore, it is probable that neither of the genetic variants results in a high impact regarding protein function, which complies with the information obtained from the impact prediction programs (PolyPhen-2, SIFT, and CADD). These programs are all in agreement with each other and classify both variants as benign.

Regarding the UCP2 protein, the information available was limited compared with the one for the LTBP4 protein. Therefore, it was not possible to have a full knowledge of this protein structure, being only known that it contains 3 solute carrier (solcar) repeats. The genetic variant rs660339 in the *UCP2* gene, causes a change from the amino acid alanine to valine at the protein position 55. This variant, present in a conserved interspecies region, results in a change from a non-essential amino acid with a nonpolar side chain (alanine) to an essential amino acid that also possesses a nonpolar side chain (valine). This change, theoretically, only causes a low functional impact since both of the amino acids belong to the same group and share similarities. Based on the information provided by UniProtKB, the amino acid change occurs in the solcar 1 repeat. The information obtained from the different impact prediction programs (PolyPhen-2, SIFT, and CADD) is in concordance with each other, since all of them classify this genetic variant as benign.

The splice variants do not possess a protein functional impact. Nonetheless, the variant rs2589156 in the *RPTOR* gene can have as a consequence a possible loss of the splice region. Therefore, the transcript that could be originated from this variant was searched for. However, no transcript where this loss is described was found, being all transcripts similar for the exons and introns positions. As the documented transcripts belong to supposedly healthy individuals, the absent of this possible transcript gives emphasis to the likely pathogenicity of this variant. The remaining variants, rs2304483 in the *SLC12A3* gene and rs10169718 in the *ARPC2* gene, were considered benign variants, which suggest that they do not affect the splice region. In compliance with this is the fact that no protein-coding transcripts were found as a consequence of these variants.

3.2. Candidate Genes

The candidate gene approach is one of the approaches being used to expand the knowledge regarding mechanisms linking the diabetic milieu to tissue damage, hoping that location and function of the identified variants will point to genes and molecular pathways that can be involved in the etiology of these conditions (158). This approach assumes a high probability of association between the development of the disease being studied and common genetic variants, based on the gene known function (40). Nonetheless, and even though valuable, it is difficult to draw firm conclusions from these studies. Some of the performed studies have often been small and have yielded p-values that are only nominally significant and cannot withstand an adjustment for multiple comparisons. In other cases, limited or no attempts were made to replicate the findings and, when those attempts were made, multiple reports were published with conflicting findings resulting either from false positives due to the marginal p-values or from false negatives due to the lack of statistical power. However, and despite these problems, interesting associations have emerged (158).

So far, only a small number of genetic variants in an also small number of candidate genes were identified (40). The complete list of the candidate genes, as well as their respective genetic variants, for European type 2 diabetic individuals, is available in Appendix D – Table D3. From those 19 candidate genes and their genetic variants, none of them was found in the common variants list that resulted from the statistical analysis of the population in this study.

The candidate genes found so far were selected on the basis of their postulated role in cellular pathways linking glucose to tissue damage, being mainly genes implicated in the pathways regulating the production of cellular toxins in response to hyperglycemia, namely, ROS, sorbitol and AGEs, and genes from pathways involved in the tissue-specific organ damage induced by these toxins (40,158). The *ADIPOQ* gene encodes for adiponectin, a cytokine exclusively produced by adipocytes that has insulin-sensitizing effects (158). Adiponectin levels are high in cases of diabetic nephropathy, but it remains unclear whether these high levels are a cause or a consequence of the disease. In the study performed it was verified that the two associated variants in this gene (rs17300539 and rs2241766) are determinants of renal risk by leading to high circulating adiponectin concentrations (189). In the gene that encodes the receptor for AGEs, the *AGER* gene, a SNP was identified as a significant predictor for diabetic nephropathy (2184A>G). Nowadays, there is an increasing amount of evidence supporting the role of AGEs, for which AGER is a receptor, in diabetic nephropathy pathogenesis (190).

Of 31 SNPs associated with diabetic nephropathy in African Americans, the SNPs rs7285167 in the *APOL2* gene, as well as rs61098917 and rs3747154 in *LIMK2* and rs4478844 in *OR2AK2*, were also found to be associated with this complication in a European American population. The *APOL2* gene encodes an apolipoprotein, while *LIMK2*, a LIM kinase, is involved in the regulation of the actin cytoskeleton, and *OR2AK2* is an odorant receptor (191). The carnosine dipeptidase genes, *CNDP1* and *CNDP2*, present two and one variants associated with diabetic nephropathy, respectively. The 5-5 leucine repeat polymorphism in *CNDP1* was associated with a significantly reduced risk of developing diabetic nephropathy, while rs2346061 also in this gene and rs7577 in the *CNDP2* gene were associated with an increased risk of developing the complication. These variants are located in the regulatory region of their respective genes, and could thereby modulate carnosinase activity. The *CNDP1* encodes a dipeptidase that hydrolyses the substrate L-carnosine (β-alanyl-L-histidine) specifically while *CNDP2* encodes a non-specific dipeptidase. Carnosine has been linked to diabetic nephropathy, since it has been described as having anti-oxidant effects since it serves as a scavenger of oxygen radicals and thus can inhibit the formation of AGEs (192,193).

The *EPO* gene encodes the angiogenic factor erythropoietin (EPO). In this gene, a SNP in the promoter region was identified and associated with elevated EPO, resulting in an

increased risk for the development of diabetic nephropathy (158,194). Furthermore, some reports suggest that iron overload might cause diabetic nephropathy. An increased accumulation of iron could affect insulin synthesis and secretion in the pancreas, as well as it could enhance oxidation of free fatty acids through accelerated production of free radicals. Moreover, it was suggested that accumulation of iron might causes kidney damage during the course of diabetes. Therefore, a study was conducted to evaluate the role of genetic variants in the hemochromatosis gene, the *HFE* gene, as a risk factor for type 2 diabetes associated diabetic nephropathy. This study concluded that rs1799945 in this gene was associated with the development of diabetic nephropathy (195). The *HMGA2* gene is ascribed as having an important role in the control of stem-cell development and proliferation, being an increase in its expression found in many benign, as well as malignant, tumors. Although the role of this gene in diabetic nephropathy or even in diabetes is still not fully understood, it was demonstrated that overexpression of *HMGA2* was associated with the formation of micropolycystic kidney, which suggested the involvement of this gene in kidney development. Furthermore, changes in *HMGA2* expression may ultimately increase the risk of developing irreversible glomerulosclerosis and tubulointerstitial fibrosis, both of which are involved in the pathogenesis of diabetic nephropathy. The performed study demonstrated that the *HMGA2* variant (rs1531343) seems to be associated with increased risk of developing nephropathy in patients with type 2 diabetes (196).

The heat-shock proteins (HSPs) are molecular chaperones, synthesized under stress conditions, involved in renal cell survival and matrix remodeling in acute and chronic renal diseases. A study performed to investigate whether gene polymorphisms in *HSPA1A* affected diabetic nephropathy susceptibility in type 2 diabetic individuals indicated that rs1008438 and rs1043618 in this gene were associated with a predisposition to diabetic nephropathy (197). An association between the *IL1RN* gene, an interleukin-1 receptor antagonist, and several inflammatory diseases has already been found, being this gene implicated in that disease inflammatory mechanism. Therefore, an association was tested between this specific gene and complications of diabetes which have an inflammatory tissue component. From this experiment, an 86 bp variable number tandem repeat polymorphism in the intron 2 of this gene was found to be linked with T2D associated renal complications, being the functional significance of that polymorphism still being

investigated (198). The *MYH9* gene encodes non-muscle myosin IIA and is expressed in glomerular podocytes and mesangial cells. Polymorphisms in this gene are strongly associated with idiopathic and HIV-associated forms of focal segmental glomerulosclerosis (FSGS) and nondiabetes-associated ESRD in African Americans, idiopathic FSGS and non-diabetic kidney disease in European Americans and non-diabetic ESRD in Hispanic Americans. However, gene polymorphisms in *MYH9* have not been evaluated for association with diabetic nephropathy and consequent ESRD in type 2 diabetic individuals in a European American population. The conducted study presented the variants rs4821480, rs4281481, rs2032487 and rs3752462 as associated with the development and progression of diabetic nephropathy to ESRD in European Americans (199).

The endothelial isoform of the nitric oxide synthase, encoded by the *NOS3* gene, is a well-documented functional candidate gene for diabetic nephropathy susceptibility due to its involvement in catalyzing the production of nitric oxide, which is involved in the regulation of the vascular tone (190,200). So, in the performed study with the objective of investigating this gene possible role in the progression of nephropathy, the genetic variant 894G>T was identified. This variant was considered a risk factor for the progression of diabetic nephropathy in type 2 diabetic individuals (200). In a study performed to evaluate an association between variants and diabetic nephropathy in individuals with type 2 diabetes, 11.152 SNPs were investigated. From those, the top ranked SNPs were rs2285372 and rs2301572 in the *PLXND1* gene, and rs1543547 in *RAET1L*. The *PLXND1* gene is expressed in human vascular endothelial cells and is required for the normal development of vasculature, while genes from the same family as *RAET1L* encode glycoproteins that are anchored to the membrane via GPI linkage (201). The *SOD1* gene encodes for a superoxide dismutase enzyme. These enzymes play a major role in detoxification of ROS and protection against oxidative stress, since they catalyze the dismutation of superoxide into oxygen and hydrogen peroxide. Associations of *SOD1* gene variants with diabetic nephropathy were reported in patients with type 1 diabetes. Therefore, in a conducted study in 2012, the association of variations in the *SOD1* gene with nephropathy in patients with type 2 diabetes was investigated. From this study resulted the variant rs1041740, which showed an association with the prevalence of incipient nephropathy (microalbuminuria) (202).

As already known, TGF-β is one of the triggers of ECM protein overproduction, being an increased TGF-β1 production been reported as associated with the development of diabetic nephropathy. Thereby, the study conducted aimed to investigate the potential role of variants in the *TGFB1* gene in microvascular diabetic complications such as diabetic nephropathy. The results of this study showed that the 869T>C variant in this gene was associated with an increased susceptibility to the development of diabetic nephropathy (203). Lastly, regarding the *UMOD* gene, several studies have identified this gene, which encodes the most common protein in human urine, to be associated with hypertension, a risk factor for diabetic nephropathy, and also with chronic kidney disease (CKD). Thus, the objective of the performed study was to examine the association of a single common variant in this gene (rs13333226) with type 2 diabetic nephropathy and kidney function. From the results obtained, this variant was associated with a decreased risk of nephropathy, being also associated with better kidney function and lower blood pressure. Nonetheless, the association with diabetic nephropathy was independent of blood pressure and kidney function (204).

3.3. Validation

The common variants described in Table 13 were validated by two different methods. First, each exome BAM file was manually verified for the positions of the genetic variants present in that same exome. This verification allows to distinguish between the real variants that were correctly identified in the variant calling step and the variants that were only identified in that step due to sequencing errors, being, therefore, an important validation step for all the genetic variants. An example of a BAM file for a wild-type, an altered homozygous and a heterozygous exome for each common variant is presented in Figure 33.

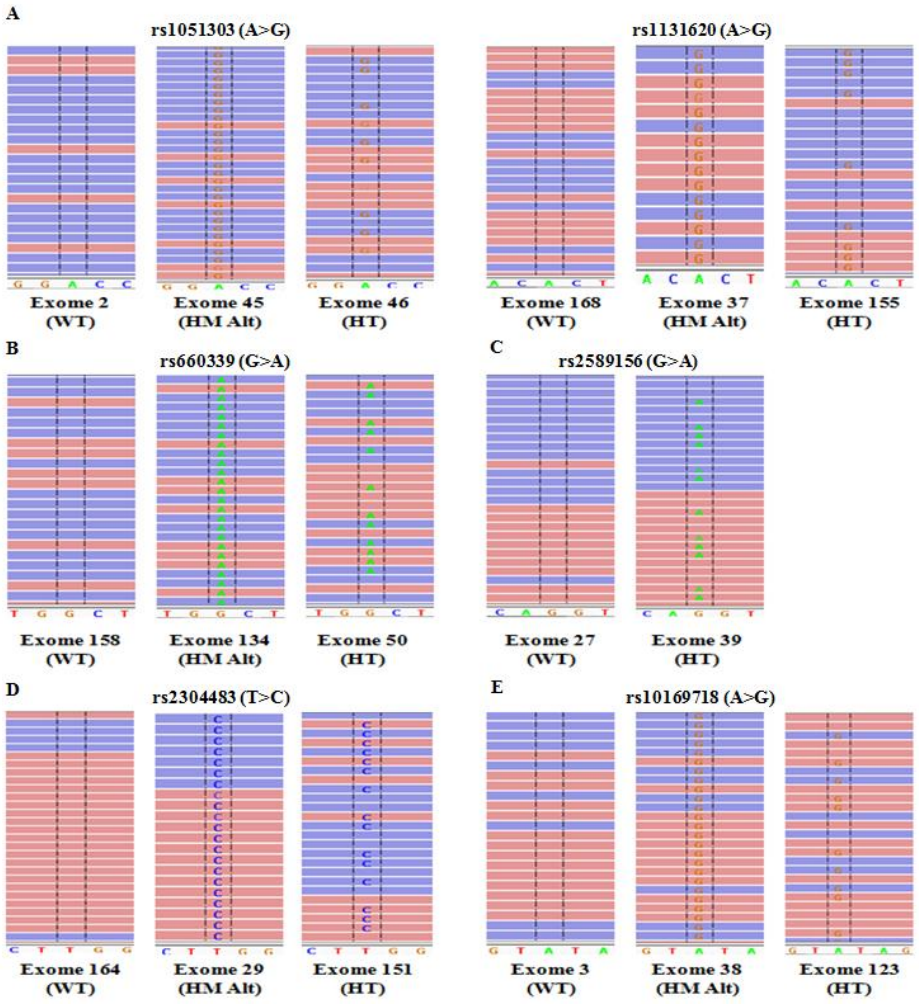

Figure 33. BAM file example for a wild-type, an altered homozygous and a heterozygous exome for each common variant. In (A) are represented the variants present in *LTBP4*, (B) refers to the variant present in *UCP2*, (C) considers the variant in *RPTOR*, (D) shows the genetic variant present in *SLC12A3* and in (E) the common variant localized in *ARPC2*. The blue reads are the reverse strands and the pink reads represent the forward strands. HM Alt: altered homozygous; HT: heterozygous; WT: wild-type.

In addition to the BAM files verification, all of the biologically relevant common variants obtained were also validated by the technique used to validate the sequencing technology. For that, 4 of the 36 exomes were genotyped by the Illumina microarray with the HumanOmniExpressExome. In this procedure, the various genotypes obtained by the Illumina microarray regarding specific variants in those 4 exomes were compared to the genotypes obtained by the Ion Proton™ Sequencer for the same specific variants in the

same exomes, being those genotypes a match. The validation by this microarray is only possible for variants with rs ID.

All the common variants were, thereby, checked and validated as real, both by the visualization of the BAM files, as well as by the comparison between the results obtained from the microarray with the ones obtained by the Ion Proton^{TM} Sequencer. Therefore, in the common variants approach, the performed validation procedures confirmed all of the results obtained by exome sequencing.

4. Rare Variants Approach

4.1. Statistical Analysis

Until now, the common variants studies conducted have had little success in implicating specific genes in specific diseases, being thought that rare genetic variants can play a key role in influencing the development of complex diseases and its related traits (104,205).

However, standard statistical methods used to test for association with single common genetic variants are underpowered for rare variants, unless sample sizes or effect sizes are very large. An alternative approach are burden tests that can assess the cumulative effects of multiple rare variants in a determined genomic region (205). Therefore, in this study, the rare genetic variants were grouped by gene using EPACTS (154), followed by a binary (cases vs. controls) gene-wise burden test. Statistically significant results, p-value ≤ 0.05, were obtained for 128 rare variant accumulating genes. Those genes are presented in Appendix E – Table E1. From the results obtained, each one of the statistically significant genes with accumulated rare variants was studied, based on the literature, to identify the most biologically relevant genes to the pathogenesis of diabetic nephropathy. From that search, 3 genes accumulating rare variants were selected as the most relevant genes. Of those, 2 genes (*STAB1* and *CUX1*) were considered protective genes, presenting genetic variants mainly present in the control group. The remaining gene (*MMP25*) was considered a risk gene, with its accumulated variants primarily present in the case group.

The annotation of the relevant genes and their respective accumulated rare variants is presented in Table 15. The software applications PolyPhen-2 (132), SIFT (134), and CADD (135) indicated the impact of the SNP variants in the coding regions, while the

programs SplicePort: An Interactive Splice Site Analysis Tool (161), HSF (162), Analyzer Splice Tool, NNSplice (163), HBond Score Web-Interface (164) and USD SplicePredictor Online Service (165) classified genetic variants in the splice regions. The results from those programs are available in Appendix E – Table E2.

Table 15. Annotation of the rare variants accumulated in genes biologically relevant to diabetic nephropathy.

	Gene	rs ID or Chr: end position	Ref. allele	Alt. allele	Type of variant	PolyPhen-2 *	SIFT **	CADD ***	MAF (cases)	MAF (controls)	p-value	Associated mechanism
Protective	*STAB1*	rs371042844	C	T	Splice	-	-	2.39	0.00	0.03	0.01	"Clearance" of AGEs
		rs41292856	A	G	Missense	Benign	Tolerated	11.49	0.00	0.03		
		rs149944392	G	A	Missense	Probably damaging	Deleterious	21.40	0.03	0.00		
		rs199636230	C	T	Splice	-	-	12.83	0.00	0.03		
		rs143836348	T	C	Splice	-	-	2.50	0.00	0.03		
		Chr 3: 52546850	G	A	Missense	Benign	Tolerated	8.11	0.00	0.03		
		Chr 3: 52548167	G	A	Missense	Possibly damaging	Tolerated	16.07	0.00	0.03		
		rs148915659	A	G	Missense	Benign	Deleterious	9.94	0.00	0.05		
		rs143242234	C	G	Missense	Possibly damaging	Tolerated	17.20	0.00	0.03		
Risk	*MMP25*	Chr 16: 3108251	T	C	Missense	Probably damaging	Deleterious	17.64	0.03	0.00	0.01	ECM degradation
		Chr 16: 3108573	C	A	Missense	Probably damaging	Deleterious	14.70	0.03	0.00		
Protective	*CUX1*	Chr 7: 101758496	A	T	Missense	Benign	Tolerated	15.94	0.00	0.03	0.04	Regulation of collagen I transcription
		rs148760130	G	A	Missense	Benign	Tolerated	16.10	0.00	0.03		

AGEs: advanced glycation end-products; Alt.: altered; CADD: combined annotation dependent depletion; Chr: chromosome; ECM: extracellular matrix; MAF: minor allele frequency; PolyPhen-2: polymorphism phenotyping v2; Ref.: reference; SIFT: sorting intolerant from tolerant.
*PolyPhen-2 (132): "benign" (≥ 0 and ≤ 0.452); "possibly damaging" (≥ 0.453 and ≤ 0.956) and "probably damaging" (≥ 0.957 and ≤ 1)
**SIFT (134): "deleterious" (≤ 0.05) and "tolerated" (> 0.05)
***CADD (135): higher scores corresponds to a higher pathogenicity (in this study a genetic variant was considered pathogenic with a score ≥ 12)

The variants rs371042844, rs199636230 and rs143836348 in the *STAB1* gene are splice region variants. For rs371042844, all of the prediction programs displayed concordant results, with the scores from each software showing little or no variation for the sequence with the variant in comparison to the reference sequence. For the rs199636230 and rs143836348, the prediction programs were not all in accordance. However, the majority of those programs exhibited scores for the sequence with the variant that were similar or close to the scores obtained for the reference sequence. Therefore, all of these variants in the *STAB1* gene can be considered benign variants, which suggests that they do not affect the splice region.

The genes accumulating rare variants found with this approach are associated with different mechanisms. The *STAB1* gene is responsible for the "clearance" of AGEs, the *MMP25* gene is involved in ECM degradation, and lastly, the *CUX1* gene presents a biological function associated with the regulation of collagen I transcription.

Glycation describes the spontaneous reaction of nucleophilic groups of biomolecules with a broad spectrum of reactive carbonyl compounds, such as glucose, ribose, glyoxal, and MG. This reaction affects nucleic acids, lipids and, most importantly, proteins. Within proteins, glycation occurs mostly at the N-terminal and in the side chains of arginine and lysine residues due to their abundance and chemical reactivity. This reaction end-products are collectively referred to as AGEs, and its formation is slow, taking place over days and weeks (206). These products, generated by nonenzymatic glycation mainly of proteins, elicit a wide variety of cellular responses, including, induction of growth factors and cytokines, adhesion molecules, chemokines and oxidant stress production (207,208). These responses are thought to contribute to the development of pathologies associated with aging, DM and Alzheimer's disease (207). In T2D, the increased formation of AGEs due to the chronic exposure to hyperglycemia can contribute to the development of complications associated with diabetes, namely diabetic nephropathy (Figure 34) (206,209,210).

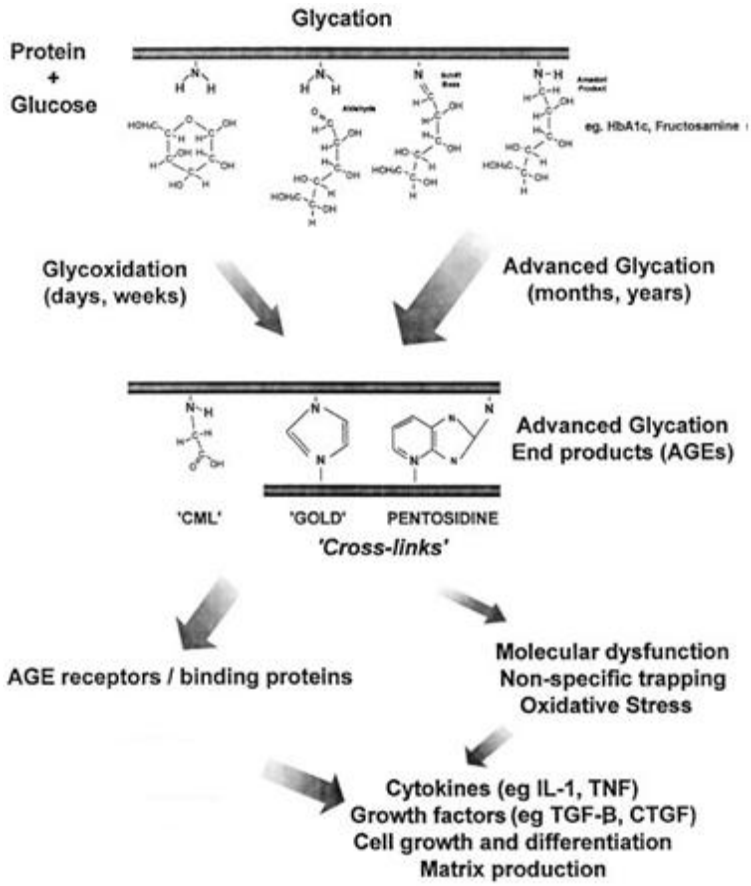

Figure 34. Formation of AGEs (adapted from (210)). AGEs: advanced glycation end-products; CML: N(6)-Carboxymethyllysine; CTGF: connective tissue growth factor; GOLD: glyoxal-lysine dimer; HbA1c: glycated hemoglobin; IL-1: interleukin-1; TGF-β: transforming growth factor β; TNF: tumor necrosis factor.

The AGEs can cause tissue damage by two main pathways, they either form cross-links that disrupts the structure and function of short and long-lived proteins and lipids, or they either interact with specific and nonspecific cell surface receptors, leading to altered intracellular events that induce oxidative stress and inflammation (208). The receptors and binding proteins through which they may exert their effects are classified as either inflammatory or "clearance" receptors. The inflammatory receptors include the advanced glycation end-product receptor (AGER, also known as RAGE), while the "clearance" receptors account for the cell surface glycoprotein cluster of differentiation 36 (CD36); the oligosaccharyl transferase complex protein 48, also known as advanced glycation end-product receptor 1 (AGE-R1); the 80K-H protein, also known as advanced glycation end-product receptor 2 (AGE-R2); galectin-3, also known as advanced glycation end-product

receptor 3 (AGE-R3); the scavenger receptor class B type I (SR-BI) and type II (SR-BII); the macrophage scavenger receptor class A type I (SR-AI) and type II (SR-AII) and also the stabilin-1 (STAB1, also known as FEEL-1) and stabilin-2 (STAB2, also known as FEEL-2) (42,206,207,209). Circulating AGE levels reflect the equilibrium between endogenous formation and catabolism, including tissue degradation and renal elimination (208). Modulation of these products receptors expression has been shown to be important in the development of diabetic nephropathy, being changes in that expression closely associated with renal impairment (209).

In diabetic nephropathy, AGEs can become accumulated in tissues, being its levels correlated with the complication severity. The mechanisms by which these products are involved in the development of diabetic nephropathy include modifications of various ECM proteins, such as collagens (non-receptor dependent mechanisms), as well as ROS generation through the interaction of tissue accumulated AGEs with AGER (receptor mediated mechanisms) (42). Because of their slow turnover, ECM proteins are highly susceptible to AGE modification, resulting in both altered structures and functions (210). The glycation of collagens yields cross-linking of these molecules, which leads to structural alterations, such as changes in packing density and surface charge, manifested by increased stiffness, reduced thermal stability and resistance to collagenase digestion (42,210). Also, interactions between matrix proteins can also be disturbed by AGE modifications. The affinity of laminin and fibronectin for type IV collagen and heparan sulfate proteoglycan is decreased after AGE modification. Furthermore, not only are AGE-modified proteins more resistant to enzymatic digestion but also AGEs accumulation in vascular tissues is associated with a reduction in the matrix degradative capacity of the kidney (210). The interaction of AGEs with their receptor, AGER, may generate ROS through stimulation of NADPH oxidase. This ultimately contributes to the induction of growth factors such as TGF-β and CTGF (42,210,211). These factors, induced via the MAPK signaling pathway, are known to mediate the overproduction and decreased degradation of several ECM proteins (211). Moreover, AGEs directly potentiates the formation of ROS through catalytic sites in their molecular structure (210). All of these factors combined, in particular, the induction of TGF-β, eventually disrupt the metabolic turnover of the ECM, causing ECM accumulation, a diabetic nephropathy hallmark (42,210). That accumulation ultimately leads to mesangial expansion and glomerular

basement membrane thickening (211). Therefore, both the enhanced formation as well as the decreased clearance of AGEs is responsible for their tissue accumulation, resulting in a higher AGE load that requires clearance and detoxification as a protective mechanism (209,211). The *STAB1* gene encodes for an endocytic "clearance" receptor for AGEs, the STAB1. The structure of this receptor is unique and unrelated to other "clearance" receptors, but its precise functions remain unclear. The STAB1 scavenges AGEs accumulated in vascular tissues in pathologies associated with aging or diabetes, playing a significant role in their elimination. Thereby, this receptor can be implicated in pathologies associated with aging or diabetes and can directly contribute to the development of diabetic vascular complications (207).

The gene *MMP25* is involved in ECM remodeling. During this remodeling, cleavage of the different ECM components is important for regulating matrix abundance, composition and structure (73). The MMPs have long been identified as critical mediators of ECM degradation and collectively they can degrade all of the ECM proteins (73,75). In diabetic nephropathy, the balance between ECM synthesis and degradation is one of the most important processes for maintaining the glomerular structural and functional integrity. Therefore, there is an increasing amount of evidence supporting an important role for MMPs in this complication development and progression, since the imbalance between ECM synthesis and degradation could lead to abnormal ECM deposition (75). The *MMP25* gene encodes for the MMP25 protein, also known as MT6-MMP, a GPI-anchored protein. This MMP is a relatively recent member of the MT-MMPs and has the capability of degrading specific substrates, such as collagen IV, gelatin and fibronectin (73,75,212). The type IV collagen and fibronectin are known ECM proteins that are accumulated in the kidney of individuals with diabetic nephropathy (44). Furthermore, in addition to those substrates, MMP25 also cleaves others MMPs, such as pro-MMP2 and pro-MMP9, activating them (73,74,75). In MMP9, a dinucleotide repeat polymorphism was shown to be protective against the development and progression of diabetic nephropathy (213).

The *CUX1* gene acts as a transcriptional modifier, being reported that this gene transcriptionally suppresses several other genes. In diabetic nephropathy, TGF-β, more specifically the isoform TGF-β1, has been characterized as a cytokine that plays a vital role in inducing the synthesis of matrix proteins, including collagen I. This growth factor shifts the balance towards the overproduction of ECM proteins, leading to scarring (214). The

gene *CUX1* encodes the CUX1 protein, a member of the homeodomain family of DNA binding proteins that acts as a transcription factor. The majority of published studies describe it as a transcriptional repressor (22,214). This gene is important for kidney development, with its abnormal expression being implicated in kidney associated diseases (22). The transcription of the *COL1A2* gene, the gene responsible for collagen I, is tightly regulated by transcription factors. The proximal promoter of *COL1A2* is under the control of a CCAAT motif that is located at −80 bp relative to the transcriptional start site (TSS) and is recognized by a protein called CCAAT binding factor (CBF)/nuclear factor (NF)-Y. Studies have showed that single nucleotide base changes on the CCAAT motif of *COL1A2* lead to a defective transcription of the type I collagen gene. The CUX1 protein has been reported to carry a CCAAT displacement activity that enables it to compete for binding with CBF/NF-Y in relevant promoter/enhancers of genes. Therefore, CUX1 suppresses type I collagen by inhibiting the normal transcription of its gene, in a dose-dependent manner. This protein physically interacts with the *COL1A2* proximal promoter, which leads to a partial competition for binding occupancy with CBF/NF-Y. As CBF/NF-Y is an activator of type I collagen gene expression, its displacement and therefore inability to associate with the promoter, results in a significant decrease in promoter activity. With the reduced promoter activity, a reduction of *COL1A2* transcriptional activity also occurs, leading to a direct down-regulation of *COL1A2* transcription. Furthermore, *CUX1* has been shown to be a TGF-β responsive gene, since the growth factor induces *CUX1* expression in a dose-dependent manner. In addition to this, TGF-β also promotes the movement of CUX1 to the nucleus, where it then becomes bioavailable to exert its inhibitory effects. In conclusion, a deregulation of the *CUX1* gene may be a key factor leading to renal abnormalities due to ECM proteins accumulation. However, the repression of collagen I by transcriptionally switching off the collagen gene cannot be permanent, because collagen I is required during normal development and is necessary for tissue remodeling (214).

Furthermore, the rare variants accumulated in the biologically relevant genes, were localized, to determine if some gene presented a rare variant “hotspot” (Figure 35).

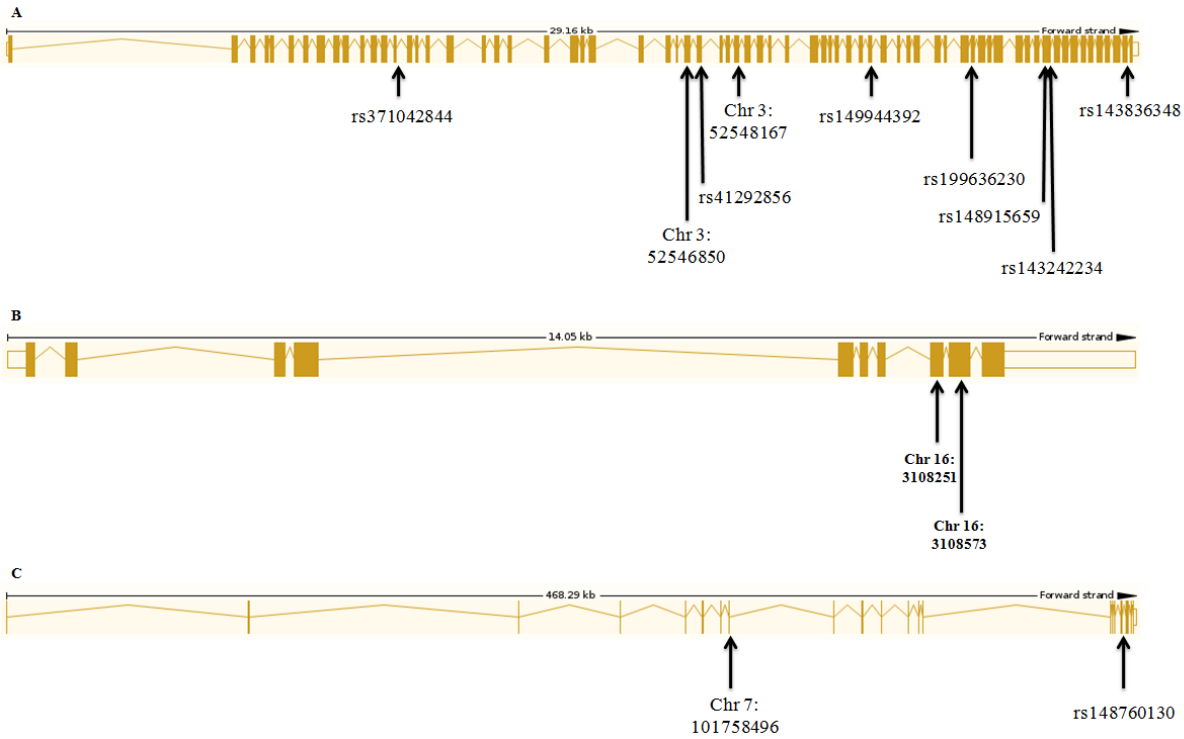

Figure 35. Localization of the rare variants accumulated in their respective genes. In (A) are the 9 variants present in *STAB1*, (B) refers to the 2 genetic variants present in *MMP25* and (C) considers the 2 genetic variants accumulated in *CUX1*. The gene sequences used were from version GRCh37 of Ensembl (139).

The *STAB1* gene is the gene with the highest number of rare variants accumulated. However, most of those variant are spread across the gene exons, not being evident the presence of a variant "hotspot". Nonetheless, the variants rs148915659 and rs143242234 share the same exon, creating a possible "hotspot". Furthermore, also the variants Chr3: 52546850 and rs41292856, even though not present in the same exon, are in neighboring exons. Nevertheless, even though the STAB1 gene presents a high number of rare variants, for the confirmation of the possible "hotspots" a greater number of variants are necessary.

Regarding the *MMP25* and *CUX1* genes, the presence of a variant "hotspot" could not be determined, since the number of rare variants present in each gene was reduced.

The missense rare variants, similarly to the common variants, also have an impact in terms of proteins, since they are variants localized in the gene exons. Therefore, the

changes in amino acids that occur as a consequence of those genetic variants are presented in Table 16.

Table 16. Functional impact of the missense rare genetic variants.

Gene	rs ID or Chr: end position	Ref. amino acid	Alt. amino acid	Change position/Protein length	Codon change
STAB1	rs41292856	Serine	Glycine	1089/2570	Agt/Ggt
	rs149944392	Glycine	Serine	1529/2570	Ggt/Agt
	Chr 3: 52546850	Glycine	Serine	1012/2570	Ggc/Agc
	Chr 3: 52548167	Alanine	Threonine	1162/2570	Gcc/Acc
	rs148915659	Threonine	Alanine	2116/2570	Act/Gct
	rs143242234	Proline	Alanine	2135/2570	Ccc/Gcc
MMP25	Chr 16: 3108251	Phenylalanine	Serine	359/562	tTc/tCc
	Chr 16: 3108573	Aspartate	Glutamate	440/562	gaC/gaA
CUX1	Chr 7: 101758496	Lysine	Isoleucine	217/676	aAa/aTa
	rs148760130	Arginine	Glutamine	558/676	cGg/cAg

Alt.: altered; Chr: chromosome; Ref.: reference.

The information available regarding the STAB1 protein structure is reduced. Therefore, it was only possible to gather that this protein is comprised of 2570 amino acids and includes 7 fasciclins I (FAS1), 16 EGF-like repeats, 2 laminin EGF-like and 1 link domain near the transmembrane region (207). The FAS1 domains are present in secreted and membrane-anchored proteins, being that most of those proteins contain multiple FAS1 domains, either in tandem or interspersed with other domains. Although the biological functions of many of these domains remain to be elucidated, it has been suggested that the FAS1 domain represents an evolutionarily ancient cell adhesion domain common to plants and eukaryotes (215). The EGF-like domain consists of approximately 40 amino acids residues in length and is characterized by a conserved arrangement of 6 cysteines residues which form disulfide bonds within the domain (Figure 36).

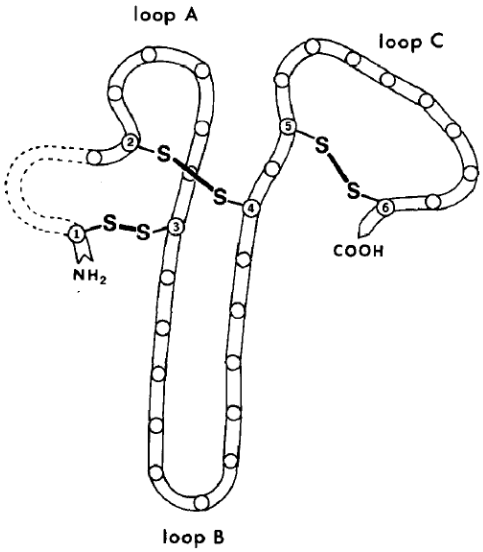

Figure 36. Structure of EGF-like domains (taken from (216)). The numbers represent the cysteine residues with the disulfide bonds.

From this domain it is known that the EGF-like repeats of various proteins are involved in receptor-ligand interactions; that a common structural folding is shared by all the EGF-like domains, based on the sequence homologies and dictated by the position of the disulfide bonds, and lastly, that this common structure forms three loops, which together provide specificity and maximum binding affinity in the proteins with this domain, being the recognition sequence located in loop B (216).

In this gene, the rare variant rs41292856 causes a change from the amino acid serine to glycine at the protein position 1089. This change switches a non-essential amino acid with a polar side chain (serine) for an also non-essential amino acid but with a nonpolar side chain (glycine), in an interspecies conserved region. Based on the information provided by UniProtKB, this change occurs in the third FAS1 domain (FAS1 3). Even though this change switches an amino acid with a polar side chain for one with a nonpolar, both of the amino acids are non-essential. Therefore, theoretically, this change does not provide a significant functional impact. In accordance to this is the information obtained from the impact prediction programs (PolyPhen-2, SIFT, and CADD), which uniformly classified this variant as benign.

The variants rs149944392 and Chr 3: 52546850, change the amino acid glycine to serine at the protein positions 1529 and 1012, respectively. These variants, present in interspecies conserved regions, result in the alteration of a non-essential amino acid with a

nonpolar side chain (glycine) to an also non-essential amino acid with a polar side chain (serine). According to UniProtKB, the variant rs149944392 is located in the twelfth EGF-like domain and the Chr 3: 52546850 variant is present in the FAS1 3 domain of the protein. Although these changes switches an amino acid with a nonpolar side chain for one with a polar one, both of the amino acids are non-essential, allowing the speculation that maybe this change would not provide a significant functional impact for both variants. However, for the variant rs149944392, located in the twelfth EGF-like domain, all of the impact prediction programs (PolyPhen-2, SIFT, and CADD) classified this variant as a pathogenic one. This could be due to the location of the variant in the protein, affecting an important domain, and not due to the amino acid change itself. Regarding the Chr 3: 52546850 variant, located in the FAS1 3 domain, the information obtained from the prediction programs (PolyPhen-2, SIFT, and CADD) is in compliance with the low impact amino acid change. All of the programs classified this variant as benign.

The variant Chr 3: 52548167 results in an exchange from the amino acid alanine for the amino acid threonine at the position 1162. This exchange occurs between a non-essential amino acid with a nonpolar side chain (alanine) and an essential amino acid with a polar side chain (threonine), in an interspecies conserved region. The information obtained in UniProtKB localizes this variant in the fourth FAS1 domain (FAS1 4). Based on the amino acid change it can be speculated that the functional impact of this variant in terms of protein function can be high. This can be justified by the fact that it is the substitution of a non-essential amino acid by an essential one, but mainly, because it is a switch between amino acids belonging to distinct groups possessing very different characteristic (nonpolar side chain amino acid switched for an amino acid with a polar one). The information from the prediction programs for this variant is not consistent, since PolyPhen-2 and CADD classified the variant as pathogenic while SIFT considers it benign.

The variant rs148915659 changes a threonine residue for an alanine residue at the protein position 2116. This switch occurs in an interspecies conserved region between an essential amino acid with a polar side chain (threonine) and a non-essential amino acid with a nonpolar side chain (alanine). Regarding protein localization in UniProtKB, the exchange takes place in the fifteenth EGF-like domain and can, theoretically, be considered a high functional impact exchange. This happens because the amino acid switch involves the substitution of an essential amino acid for the organism to a non-essential one.

Furthermore, also the groups to which each amino acid belongs are opposed ones (polar vs nonpolar). Once more, the prediction programs are not in accordance, being the variant considered benign in PolyPhen-2 and CADD, whereas SIFT classifies it as pathogenic, more specific deleterious.

Lastly, the variant rs143242234 leads to the substitution of the proline amino acid for alanine at the protein position 2135, an interspecies conserved region. This substitution occurs between two non-essential amino acids that present nonpolar side chains and, according to UniProtKB, is localized in the sixteenth EGF-like domain. Due to the similarity between both amino acids regarding both being non-essential to the organism, as well as belonging to the same group and, therefore, sharing the same characteristics, this variant functional impact would, theoretically, be considered low. However, the PolyPhen-2 and CADD programs considered this variant pathogenic, being SIFT the only one that considers it as benign. This could be justified by the location of the variant in the protein, maybe affecting an important domain, and not due to the amino acid change itself.

The MMPs are composed of several shared functional domains, such as the signal peptide domain, the propeptide domain, the catalytic domain and the hemopexin-like domain (73,217). The signal peptide domain, present in the N-terminal of the protein, is the domain required for the secretion of MMPs. The propeptide domain contains a cysteine (Cys)-switch motif proline-arginine-cysteine-glycine-X-proline-aspartate (PRCGXPD), where X is any amino acid. Regarding the catalytic domain, this domain has proteolytic activity, in addition to containing a zinc-binding motif. The cysteine residue in this motif interacts with the zinc ion, keeping pro-MMPs inactive until the propeptide domain is removed. In the C-terminal, the hemopexin-like domain is present, being this domain present in almost all MMPs. This domain facilitates protein-protein or lipid-proteins interactions and is involved in substrate specificity and in the non-proteolytic functions of MMPs (73). The MT-MMPs share this same basic domain organization of most MMPs, with the exception of the anchoring region (217). This specific type of MMPs is anchored to the cell surface by either a transmembrane domain followed by a short cytoplasmic tail or a GPI sequence (73). The MMP25, a MT-MMP GPI-anchored protein, as well as other MT-MMPs, regardless of being transmembrane or GPI-anchored proteins, also contain a arginine-X-arginine/lysine-arginine (RXR/KR) motif, where X is any amino acid, at the end of the propeptide domain (217). This motif serves as a recognition site for pro-

convertases such as furin, which cleaves the propeptide of inactive precursors, consequently activating the MT-MMP zymogen and releasing functional proteins (73,217). In Figure 37 the structure of the MMP25 protein is presented.

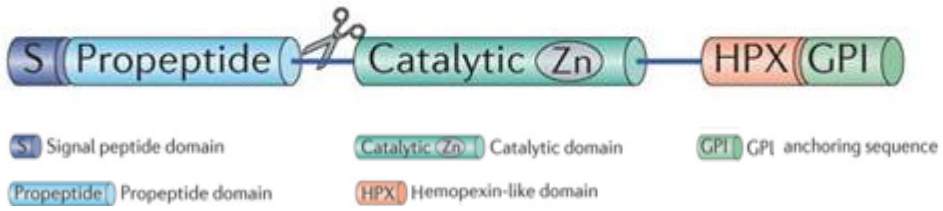

Figure 37. Structure of the MMP25 protein, a MT-MMP GPI-anchored protein (adapted from (73)). GPI: glycosylphosphatidylinositol; Zn: zinc ion.

The rare variants accumulated in the *MMP25* gene that affect this protein are Chr 16: 3108251 and Chr 16: 3108573. In Chr 16: 3108251, the change occurs between a phenylalanine residue and a serine one in the position 359. This means that an essential amino acid with a nonpolar and aromatic side chain (phenylalanine) is substituted for a non-essential amino acid with a polar side chain (serine). Regarding the Chr 16: 3108573 variant, the aspartate is the amino acid switched for glutamate, in the protein position 440. Therefore, a non-essential amino acid with a negatively charged side chain (aspartate) is substituted for an also non-essential amino acid with a negatively charged side chain (glutamate). Furthermore, both of these switches occur in interspecies conserved regions and, accordingly to the information obtained by UniProtKB, both rare variant affect the hemopexin-like domain. Theoretically, the first variant, Chr 16: 3108251, would present a high functional impact, since phenylalanine is an amino acid essential to the organism and presents a cyclic group, therefore occupying more space in the protein structure, and is replaced by serine, a non-essential amino acid with a very different set of characteristics and a plain side chain. In addition to this, the affected domain is also a crucial one, as mentioned above, for MMPs to perform their function correctly. Moreover, agreeing with the high functional impact theory are all of the impact prediction programs since all of them, PolyPhen-2, SIFT, and CADD, classified this variant as a pathogenic one. The remaining variant, Chr 16: 3108573, would possibly present a low functional impact. This is because the amino acid exchange caused by this variant occurs between two non-essential amino acids presenting the same type of side chain and, therefore, being very similar with each other. However, all of the prediction programs (PolyPhen-2, SIFT, and

CADD) classified this variant as pathogenic. This could be due to the fact that, even though similar among them, the switch occurs in the hemopexin-like domain and consequently can affect the function of the MMPs.

The CUX1 protein is an evolutionarily conserved protein that contains four DNA-binding domains (214). The CUX1 protein structural organization consists of an autoinhibitory domain present in the N-terminal, followed by a coiled-coil, three Cut repeats and a homeodomain, and in the C-terminal region, two active repression domains (218,219). Autoinhibitory domains are regions within the proteins that negatively regulate the function of other domains via intramolecular interactions. Autoinhibition is a potent regulatory mechanism that provides tight "on-site" repression and generates valuable clues to how a protein is regulated within a biological context (220). The coiled-coil is a folding motif with its most characteristic feature being a heptad repeat pattern of primarily nonpolar residues that constitute the oligomer interface. The architecture of this highly versatile folding motif determines its oligomerization state, rigidity and ability to function as a molecular recognition system (221). The Cut repeats are highly conserved 73 amino acid motifs in mammals that share from 52% to 63% amino acid identity with each other. Their degree of conservation suggests that they may carry an important biological function, and indeed, it has been shown that they bind specifically to DNA, functioning as DNA-binding domains. These domains exhibit overlapping but distinct sequence specificities, possessing a relaxed DNA-binding specificity and being able to recognize a wide range of sequences. Following these Cut repeats is the homeodomain. This domain is a 61 amino acid DNA-binding domain that harbors a histidine at the ninth amino acid of the third helix. That amino acid is responsible for determining the specificity of binding to the two bases following the domain TAAT core. Also, alterations within the homeodomain affect the DNA-binding specificity. Furthermore, even though Cut repeats can bind to DNA with high affinity on their own, the highest DNA-binding affinity and specificity is achieved when the protein homeodomain binds to DNA in conjunction with the adjacent DNA-binding domains (218). The details of the repression domains have not been defined yet. However, it is known that active transcriptional repressor proteins such as CUX1, in contrast to passive repressors, include independent repression domains and inhibit the initiation of transcription directly via the actions of these domains (222). The structure of the CUX1 protein is presented in Figure 38.

Figure 38. Structure of the CUX1 protein (adapted from (219)). CC: coiled-coil; CR1: Cut repeat 1; CR2: Cut repeat 2; CR3: Cut repeat 3; HD: homeodomain; In: autoinhibitory domain; R1: repression domain 1; R2: repression domain 2.

The rare variant Chr 7: 101758496 in the CUX1 protein cause a change from the lysine amino acid to an isoleucine, in the position 217. This alteration results in the substitution of an essential amino acid with a positively charged side chain (lysine) for an also essential amino acid, but with a nonpolar side chain (isoleucine), in an interspecies conserved region of the protein. Based on the information present in UniProtKB, this exchange is located in the coiled-coil, and this change possible functional impact can be both high or low. The high functional impact hypothesis can be justified by the different side chain characteristics of both amino acids, even though they are both essential to the organism. The low impact hypothesis can be justified by the fact that the switch leads to the substitution of a positively charged amino acid for a nonpolar one, and the coiled-coil is composed of a heptad repeat pattern of mainly nonpolar residues. Therefore, this change would not result in a significant alteration of the coiled-coil composition. The prediction programs are also not in compliance regarding the impact of this variant, with PolyPhen-2 and SIFT classifying it as a benign variant, and CADD considering it a pathogenic one.

The variant rs148760130, leads to an exchange of a non-essential amino acid with positively charges side chain (arginine) for an also non-essential amino acid with a polar side chain (glutamine) in the interspecies conserved position, 558. In UniProtKB there is no information available regarding this position protein domain localization, being only possible to affirm that this position is located in a protein region following the coiled-coil. However, it can still be speculated that this variant would present a high functional impact, since, although they are both non-essential amino acids, there is a significant difference between their side chain characteristics. The impact prediction programs for this variant are once more not in agreement. PolyPhen-2 and SIFT classified this variant as benign, while CADD evaluated it as a pathogenic variant.

The splice variants rs371042844, rs199636230 and rs143836348 in the *STAB1* gene do not possess a protein functional impact. These variants, according to the splice prediction programs were considered benign, which suggest that they do not affect the splice region.

4.2. Validation

To validate the genes obtained in the statistical analysis, several validation methods were applied to each of the genes accumulating rare variants (Table 17).

Table 17. Validation procedures for each of the rare variants accumulated in the genes relevant to diabetic nephropathy.

Gene	rs ID or Chr: end position	BAM files	MAF (PT) n=19	Illumina microarray	ASO-PCR	Sanger sequencing
STAB1	rs371042844	√	√		√	
	rs41292856	√	√		√	
	rs149944392	√	√		√	
	rs199636230	√	√		√	
	rs143836348	√	√		√	
	Chr 3: 52546850	√	√		√	
	Chr 3: 52548167	√	√		√	
	rs148915659	√	√	√		
	rs143242234	√	√	√		
MMP25	Chr 16: 3108251	√	√		√	√
	Chr 16: 3108573	√	√		√	
CUX1	Chr 7: 101758496	√	√		√	
	rs148760130	√	√	√		

ASO-PCR: allele-specific oligonucleotide polymerase chain reaction; BAM file: binary alignment/map file; Chr: chromosome; MAF: minor allele frequency; PT: Portugal.

The first variant validation method was performed by verifying each exome BAM file manually for the positions of the rare genetic variants present in that same exome. As mentioned in the common variants approach, this verification is an important step for all genetic variants, since it allows to distinguish between the real variants that were correctly identified in the variant calling step and the variants that were only identified in that step due to sequencing errors. An example of a BAM file for a wild-type and a heterozygous exome for each rare variant accumulated in the respective gene is presented in Figure 39.

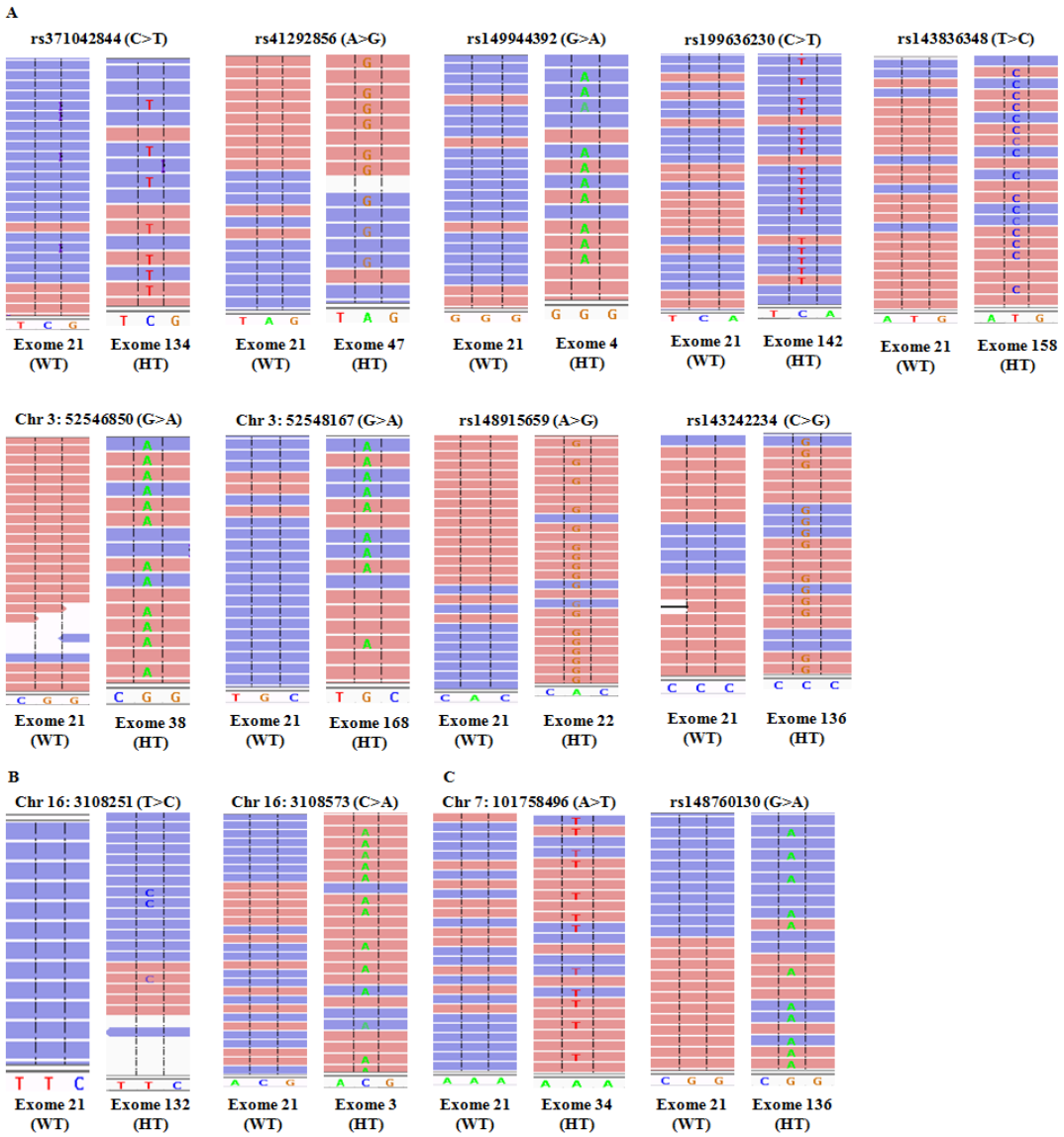

Figure 39. BAM file example for a wild-type and a heterozygous exome for each rare variant accumulated in the respective gene. In (A) the 9 variants present in *STAB1* are reprensented, (B) refers to the 2 rare variants present in *MMP25* and (C) considers the 2 variants of *CUX1*. The blue reads are the reverse strands and the pink reads represent the forward strands. HT: heterozygous; WT: wild-type.

All of the rare variants were checked and validated as real, except for the variant Chr 16: 3108251. The BAM file visualization of the heterozygous exome for this variant raised some doubts regarding the genotype veracity due to the reduced number of reads in the forward strand (pink reads) and the fact that, in that reduced number, only one of the reads presented the altered allele.

The second validation method was the determination of each rare variant MAF in the exomes of 19 healthy (without T2D and without diabetic nephropathy) Portuguese individuals unrelated to this study. This validation procedure allowed the determination of the variants frequency in the Portuguese population and therefore, the verification of the variants rarity, as well as their association with diabetic nephropathy. None of the rare variants accumulated in the genes obtained by the statistical analysis was present in any of the 19 healthy individuals exomes. Therefore, the absence of the variants in the tested healthy population confirms that the variants are indeed rare. Moreover, the fact that they are only present in type 2 diabetic individuals, regardless of whether they present or not diabetic nephropathy, but are not present in healthy individuals, gives emphasis to the association between those variants and an increased risk for the development or progression of the diabetic complication, as well as protection against it.

Furthermore, 3 rare variants (2 belonging to *STAB1* and 1 to *CUX1*) were also validated by the Illumina microarray with the HumanOmniExpressExome. This microarray genotyped 4 of the 36 exomes of individuals participating in this study, and the genotypes obtained for specific variants in those 4 exomes were then compared to the genotypes obtained for the same specific variants in the same exomes using the Ion Proton™ Sequencer. The genotypes obtained from both techniques were a match, being the tested rare variants validated as real. This comparison also allowed the sequencing technology validation, with the results obtained by exome sequencing being confirmed for the several tested variants. This microarray can only validate variants with rs ID.

The rare variants that were not validated by this latter method were subjected to ASO-PCR (10 rare variants, 7 in *STAB1*, 2 in *MMP25* and 1 in *CUX1*). The genetic variants were validated for a heterozygous and a wild-type exome for each one of the rare variants accumulating in their respective gene. Moreover, negative controls were performed for all reactions. The heterozygous and wild-type exomes used in each rare variant are shown in Table 18 and the results for each variant and the respective negative controls in Figure 40.

Table 18. Heterozygous and wild-type exomes used in ASO-PCR for each of the rare variants accumulated in the genes relevant to diabetic nephropathy.

Gene	rs ID or Chr: end position	Heterozygous exome	Wild-type exome
STAB1	rs371042844	134	21
	rs41292856	47	
	rs149944392	4	
	rs199636230	142	
	rs143836348	158	
	Chr 3: 52546850	38	
	Chr 3: 52548167	168	
MMP25	Chr 16: 3108251	132	
	Chr 16: 3108573	3	
CUX1	Chr 7: 101758496	34	

Chr: chromosome.

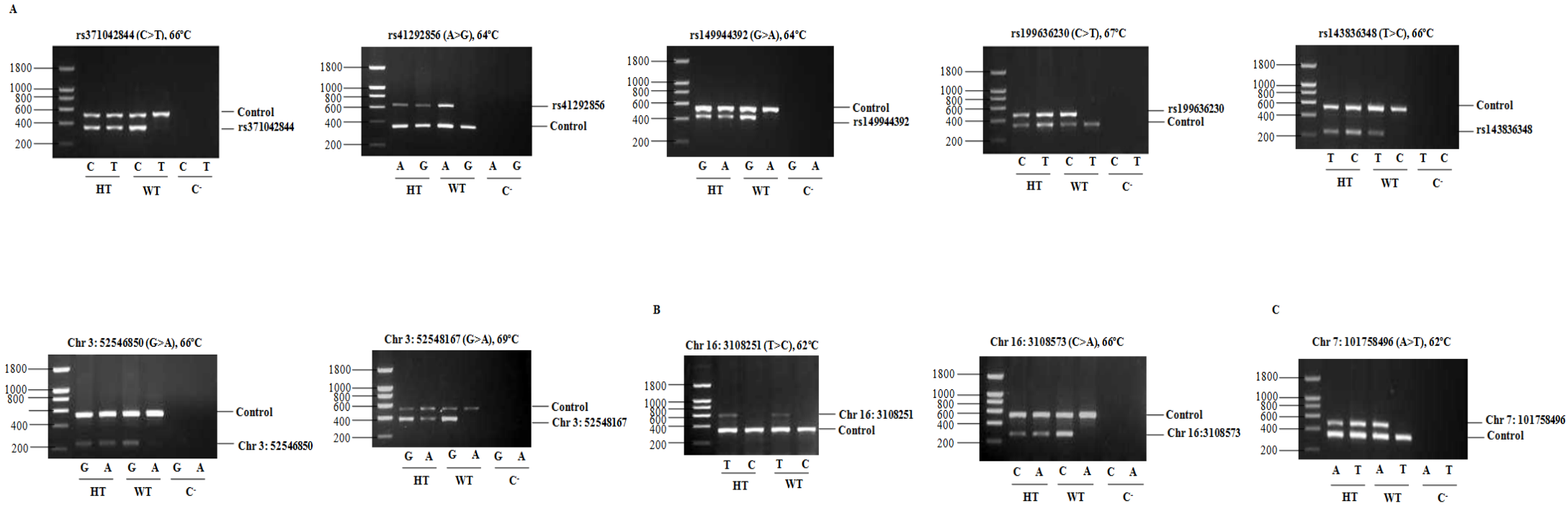

Figure 40. Electrophoresis on 1.5% (w/v) agarose gels for the rare variants accumulated in (A) *STAB1*, (B) *MMP25* and (C) *CUX1*. The molecular marker used was NZYDNA Ladder I, being its band size, in base-pairs, represented on the left side of each gel. Above the gels the genetic variant, as well as the optimized temperature for the ASO-PCR of each variant are presented and, in the right side of the gels the label for each fragment present can be seen. The alleles, as well as the exome genotype, are presented below each agarose gel. C^-: negative control; HT: heterozygous; WT: wild-type.

By the visualization of the electrophoresis gels obtained for the different rare variants, it was possible to validate the variants as real, being therefore confirmed the exome sequencing results. The exception in this validation procedure was the variant Chr 16: 3108251 in the *MMP25* gene. The results from the exome sequencing showed that exome 132 had a heterozygous genotype in this variant position. However, at the BAM file verification, this genotype had already raised some doubts regarding its veracity. Nonetheless, the variant was still subjected to ASO-PCR. Three different sets of primers were used to perform this variant ASO-PCR validation, but still, all of those sets originated results identical to the one presented in Figure 40 (B – first gel). Those results were consistent with a wild-type genotype and not with a heterozygous genotype. Therefore, to determine with certainty the exome real genotype, Sanger sequencing was used.

Sanger sequencing was performed for the exome 132, which was supposedly in heterozygosity, and for the exome 21, which presented a wild-type genotype. The objective of this procedure was to confirm exome 132 genotype. The Sanger sequencing results for both the exome 132 and 21 in the rare variant position are in Figure 41.

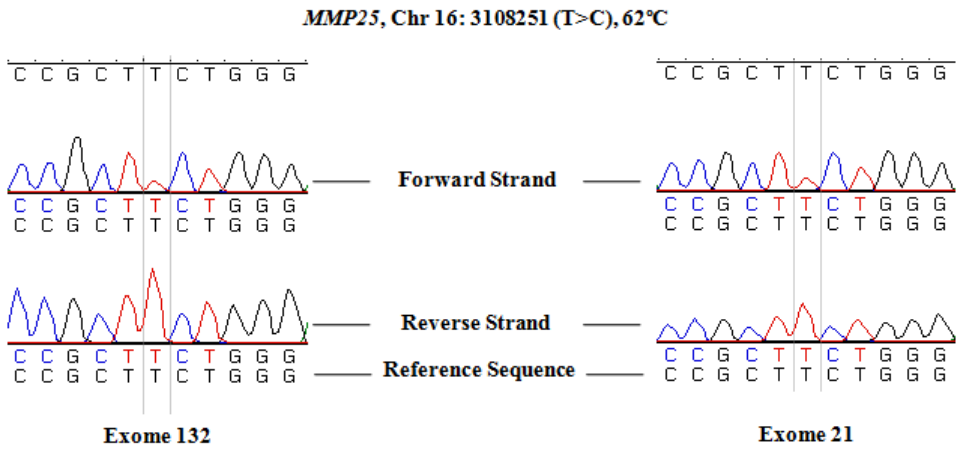

Figure 41. Sanger sequencing results for exome 132 and exome 21. The exome 132 was supposedly heterozygous (TC) while the exome 21 was a wild-type (TT).

The results from the Sanger sequencing confirmed that the exome 132 genotype, obtained from the exome sequencing, in the variant Chr 16: 3108251 position was indeed incorrect. Thus, as it can be seen in Figure 35, exome 132 presents a wild-type genotype

instead of a heterozygous genotype as originally thought. Furthermore, Sanger sequencing also allowed to confirm that exome 21 was definitively a wild-type in this position.

Exome 132 was the only exome of the 36 which supposedly presented the variant, being all others wild-types in this position. However, as the Sanger sequencing confirmed, this exome do not exhibit the variant in reality, and therefore, this variant cannot be a part of the study. As the *MMP25* gene only accumulated two rare variants in the population used for this work, and taken into account that this variant can no longer be a part of the study, the *MMP25* gene no longer accumulates rare variants. Therefore, this gene cannot be considered as a valid molecular marker for the development and progression of diabetic nephropathy in this work. This situation shows how important it is to verify manually each variant in the BAM files, as well as to perform proper validation procedures since exome sequencing still presents associated error rates.

CHAPTER 4| Conclusion

Along with the increase in T2D prevalence, it is expected that the prevalence of the complications associated with this condition also rise. Diabetic nephropathy, as well as other complications, is still poorly understood and often caught in a late stage. The identification of molecular markers for diabetic nephropathy, in addition to the already known risk factors, could contribute to a better understanding of this complication pathophysiology and also allow an early diagnosis as well as an adequate treatment. Although there is no definitive cure solution, diabetic nephropathy early diagnosis and early treatment are essential to delay the progression of this complication.

In this study, the exomes of 36 Portuguese individuals with diagnosed T2D were sequenced. From those individuals, 19 did not present diabetic nephropathy, being included in the control group, while the 17 individuals that presented the diabetic complication formed the case group. Based on the genetic differences between these groups, candidate common genetic variants as well as genes accumulating rare genetic variants that could explain the increased or reduced risk associated with the development and progression of diabetic nephropathy were identified.

In the search for common variants in the study population, 6 statistically significant (p-value ≤ 0.05) common variants present in 5 different genes were considered as the most biologically relevant ones to the pathogenesis of diabetic nephropathy. Those variants are rs1051303 and rs1131620 in the *LTBP4* gene, rs660339 in *UCP2*, rs2589156 in *RPTOR*, rs2304483 in *SLC12A3* and lastly rs10169718 present in the *ARPC2* gene. In all of these variants, the criteria used for their relevance determination was literature review, being all of the variants associated with the development or progression of diabetic nephropathy based in what is postulated in the literature about their genes biological function. The mechanisms in which these variants genes are involved include regulation of TGF-β release and activation, oxidative stress, the mTOR signaling pathway as well the RAAS pathway, systemic hypertension and the polymerization of actin.

In the rare variants approach, 2 rare variants accumulating genes, with statistical significant (p-value ≤ 0.05), were identified as the ones with a greater potential to be relevant to the onset of diabetic nephropathy. The identified genes were *STAB1* gene with 9 accumulated rare variants and the *CUX1* gene accumulating 2 genetic variants. Similarly to the proceedings in the identification of common variants, in this approach, also the relevance of the genes was assessed by literature review of the genes biological function.

These genes can be associated with both the development and progression of nephropathy, being involved in mechanisms such as the "clearance" of AGEs and the transcriptional regulation of collagen I, an ECM protein known to be accumulated in the mesangium of individuals with diabetic nephropathy. Based on the mechanisms, it is possible to verify that the rare variant accumulating genes are not involved in a specific mechanism, but rather spread across different ones. This is consistent with the description of rare variants, since they are defined as small effect variants with a strong cumulative biological impact. Therefore, subtle alterations throughout different mechanisms may have a cumulative effect responsible for the increased risk of developing diabetic nephropathy or its progression, as well as increased protection against it.

Furthermore, all the mechanisms, regardless of being associated with the common variants or the genes accumulating rare variants, seem to be implicated in ECM accumulation or podocyte disturbances (dysfunction and/or loss), both diabetic nephropathy hallmarks. With this, it can be hypothesized that treatments that target multiple mechanisms may indeed be more successful than those that target a specific one alone. Addittionally, the results obtained from both approachs show that most of the identified variants, as well as genes accumulating variants, confer protection against diabetic nephropathy.

Regarding the limitations of this study, there are two of them. The number of studied exomes was reduced, being that a higher number of exomes under study translates into a more reliable statistical analysis. The other limitation was related to the groups used to perform this study. In a case-control study, well-defined groups with similar characteristics are the key to ensure that any genetic difference between the groups is associated with the disease under study and not due to biased sampling. However, the groups in this study were heterogeneous among them for the covariate age, resulting in a statistically significant difference (p-value ≤ 0.05) between the control and case groups for this covariate. Nonetheless, the statistical analyses performed were adjusted for all the covariates, including age.

As future work, it is proposed, initially, the extension of this study to a larger population in order to confirm the results obtained. Furthermore, there is also a need to further explore the functional impact of the common variants and rare variants accumulated in the genes associated with diabetic nephropathy. For that, the protein structure should be studied, *in*

silico, to verify changes in protein folding and functional studies, *in vitro*, should be performed to prove any protein function changes.

This work was only an initial study concerning the identification of molecular markers possibly associated with the development and progression of diabetic nephropathy (risk genetic variants), as well as protection against it (protective genetic variants). However, the results obtained by this study can contribute to a deeper understanding of the genetic mechanisms associated with this diabetic complication.

CHAPTER 5| References

1. Lin Y, Sun Z. Current views on type 2 diabetes. J Endocrinol. 2010;204(1):1-11.
2. Karthikeyan R, Marimuthu G, Spence DW, Pandi-Perumal SR, BaHammam AS, Brown GM, Cardinali DP. Should we listen to our clock to prevent type 2 diabetes mellitus? Diabetes Res Clin Pract. 2014;106(2):182-190.
3. Organization WH. Definition, diagnosis and classification of diabetes mellitus and its complications. Report of a WHO consultation, part 1: diagnosis and classification of diabetes mellitus. Geneva: World Health Organization, 1999.
4. Stumvoll M, Goldstein BJ, van Haeften TW. Type 2 diabetes: principles of pathogenesis and therapy. Lancet. 2005;365(9467):1333-1346.
5. Guariguata L, Whiting DR, Hambleton I, Beagley J, Linnenkamp U, Shaw JE. Global estimates of diabetes prevalence for 2013 and projections for 2035. Diabetes Res Clin Pract. 2014;103(2):137-149.
6. Stumvoll M, Goldstein BJ, van Haeften TW. Type 2 diabetes: pathogenesis and treatment. Lancet. 2008;371(9631):2153-2156.
7. Lyssenko V, Almgren P, Anevski D, Orho-Melander M, Sjogren M, Saloranta C, Tuomi T, Groop L. Genetic prediction of future type 2 diabetes. PLoS Med. 2005;2(12):1299-1308.
8. Zimmet P, Arblaster M, Thoma K. The effect of westernization on native populations. Studies on a Micronesian community with a high diabetes prevalence. Aust N Z J Med. 1978;8(2):141-146.
9. Brunetti A, Chiefari E, Foti D. Recent advances in the molecular genetics of type 2 diabetes mellitus. World J Diabetes. 2014;5(2):128-140.
10. Das SK, Elbein SC. The Genetic Basis of Type 2 Diabetes. Cellscience. 2006;2(4):100-131.
11. Ferrannini E, Mari A. beta-Cell function in type 2 diabetes. Metabolism. 2014;63(10):1217-1227.
12. Rorsman P, Braun M. Regulation of insulin secretion in human pancreatic islets. Annu Rev Physiol. 2013;75:155-179.
13. Sam AH, Sleeth ML, Thomas EL, Ismail NA, Mat Daud N, Chambers E, Shojaee-Moradie F, Umpleby AM, Goldstone AP, Le Roux CW, Bech P, Busbridge M, Laurie R, Cuthbertson DJ, Buckley A, Ghatei MA, Bloom SR, Frost GS, Bell JD, Murphy KG.

Circulating pancreatic polypeptide concentrations predict visceral and liver fat content. J Clin Endocrinol Metab. 2014;100(3):1048-1052.

14. Fu Z, Gilbert ER, Liu D. Regulation of insulin synthesis and secretion and pancreatic Beta-cell dysfunction in diabetes. Curr Diabetes Rev. 2013;9(1):25-53.

15. DeFronzo RA. Pathogenesis of type 2 diabetes mellitus. Med Clin North Am. 2004;88(4):787-835.

16. Pierce M, Keen H, Bradley C. Risk of diabetes in offspring of parents with non-insulin-dependent diabetes. Diabet Med. 1995;12(1):6-13.

17. Groop L, Forsblom C, Lehtovirta M, Tuomi T, Karanko S, Nissen M, Ehrnstrom BO, Forsen B, Isomaa B, Snickars B, Taskinen MR. Metabolic consequences of a family history of NIDDM (The Botnia Study) - Evidence for sex-specific parental effects. Diabetes. 1996;45(11):1585-1593.

18. McCarthy MI. Genomics, type 2 diabetes, and obesity. N Engl J Med. 2010;363(24):2339-2350.

19. Lohmueller KE, Sparso T, Li Q, Andersson E, Korneliussen T, Albrechtsen A, Banasik K, Grarup N, Hallgrimsdottir I, Kiil K, Kilpelainen TO, Krarup NT, Pers TH, Sanchez G, Hu Y, Degiorgio M, Jorgensen T, Sandbaek A, Lauritzen T, Brunak S, Kristiansen K, Li Y, Hansen T, Wang J, Nielsen R, Pedersen O. Whole-exome sequencing of 2,000 Danish individuals and the role of rare coding variants in type 2 diabetes. Am J Hum Genet. 2013;93(6):1072-1086.

20. Lyssenko V, Laakso M. Genetic screening for the risk of type 2 diabetes: worthless or valuable? Diabetes Care. 2013;36 Suppl 2:S120-126.

21. Pinney SE, Simmons RA. Epigenetic mechanisms in the development of type 2 diabetes. Trends Endocrinol Metab. 2010;21(4):223-229.

22. Smyth LJ, McKay GJ, Maxwell AP, McKnight AJ. DNA hypermethylation and DNA hypomethylation is present at different loci in chronic kidney disease. Epigenetics. 2014;9(3):366-376.

23. Voight BF, Scott LJ, Steinthorsdottir V, Morris AP, Dina C, Welch RP, Zeggini E, Huth C, Aulchenko YS, Thorleifsson G, McCulloch LJ, Ferreira T, Grallert H, Amin N, Wu G, Willer CJ, Raychaudhuri S, McCarroll SA, Langenberg C, Hofmann OM, Dupuis J, Qi L, Segre AV, van Hoek M, Navarro P, Ardlie K, Balkau B, Benediktsson R, Bennett AJ, Blagieva R, Boerwinkle E, Bonnycastle LL, Bengtsson Bostrom K, Bravenboer B,

Bumpstead S, Burtt NP, Charpentier G, Chines PS, Cornelis M, Couper DJ, Crawford G, Doney AS, Elliott KS, Elliott AL, Erdos MR, Fox CS, Franklin CS, Ganser M, Gieger C, Grarup N, Green T, Griffin S, Groves CJ, Guiducci C, Hadjadj S, Hassanali N, Herder C, Isomaa B, Jackson AU, Johnson PR, Jorgensen T, Kao WH, Klopp N, Kong A, Kraft P, Kuusisto J, Lauritzen T, Li M, Lieverse A, Lindgren CM, Lyssenko V, Marre M, Meitinger T, Midthjell K, Morken MA, Narisu N, Nilsson P, Owen KR, Payne F, Perry JR, Petersen AK, Platou C, Proenca C, Prokopenko I, Rathmann W, Rayner NW, Robertson NR, Rocheleau G, Roden M, Sampson MJ, Saxena R, Shields BM, Shrader P, Sigurdsson G, Sparso T, Strassburger K, Stringham HM, Sun Q, Swift AJ, Thorand B, Tichet J, Tuomi T, van Dam RM, van Haeften TW, van Herpt T, van Vliet-Ostaptchouk JV, Walters GB, Weedon MN, Wijmenga C, Witteman J, Bergman RN, Cauchi S, Collins FS, Gloyn AL, Gyllensten U, Hansen T, Hide WA, Hitman GA, Hofman A, Hunter DJ, Hveem K, Laakso M, Mohlke KL, Morris AD, Palmer CN, Pramstaller PP, Rudan I, Sijbrands E, Stein LD, Tuomilehto J, Uitterlinden A, Walker M, Wareham NJ, Watanabe RM, Abecasis GR, Boehm BO, Campbell H, Daly MJ, Hattersley AT, Hu FB, Meigs JB, Pankow JS, Pedersen O, Wichmann HE, Barroso I, Florez JC, Frayling TM, Groop L, Sladek R, Thorsteinsdottir U, Wilson JF, Illig T, Froguel P, van Duijn CM, Stefansson K, Altshuler D, Boehnke M, McCarthy MI. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat Genet. 2010;42(7):579-589.

24. Tamayo T, Rosenbauer J, Wild SH, Spijkerman AM, Baan C, Forouhi NG, Herder C, Rathmann W. Diabetes in Europe: an update. Diabetes Res Clin Pract. 2014;103(2):206-217.

25. Cerf ME. Beta cell dysfunction and insulin resistance. Front Endocrinol (Lausanne). 2013;4:1-12.

26. Groop LC, Ferrannini E. Insulin Action and Substrate Competition. Bailliere Clin Endoc. 1993;7(4):1007-1032.

27. Bonnard C, Durand A, Peyrol S, Chanseaume E, Chauvin MA, Morio B, Vidal H, Rieusset J. Mitochondrial dysfunction results from oxidative stress in the skeletal muscle of diet-induced insulin-resistant mice. J Clin Invest. 2008;118(2):789-800.

28. Tateya S, Kim F, Tamori Y. Recent advances in obesity-induced inflammation and insulin resistance. Front Endocrinol (Lausanne). 2013;4:1-14.

29. Jewell JL, Oh E, Thurmond DC. Exocytosis mechanisms underlying insulin release and glucose uptake: conserved roles for Munc18c and syntaxin 4. Am J Physiol Regul Integr Comp Physiol. 2010;298(3):R517-531.

30. Polonsky KS, Sturis J, Van Cauter E. Temporal profiles and clinical significance of pulsatile insulin secretion. Horm Res. 1998;49(3-4):178-184.

31. Porksen N. The in vivo regulation of pulsatile insulin secretion. Diabetologia. 2002;45(1):3-20.

32. Perley MJ, Kipnis DM. Plasma Insulin Responses to Oral and Intravenous Glucose - Studies in Normal and Diabetic Subjects. Journal of Clinical Investigation. 1967;46(12):1954-1962.

33. Lillioja S, Mott DM, Howard BV, Bennett PH, Ykijarvinen H, Freymond D, Nyomba BL, Zurlo F, Swinburn B, Bogardus C. Impaired Glucose-Tolerance as a Disorder of Insulin Action - Longitudinal and Cross-Sectional Studies in Pima-Indians. New Engl J Med. 1988;318(19):1217-1225.

34. Boden G. Free fatty acids-the link between obesity and insulin resistance. Endocr Pract. 2001;7(1):44-51.

35. Rahier J, Guiot Y, Goebbels RM, Sempoux C, Henquin JC. Pancreatic beta-cell mass in European subjects with type 2 diabetes. Diabetes Obes Metab. 2008;10 Suppl 4:32-42.

36. Butler AE, Janson J, Bonner-Weir S, Ritzel R, Rizza RA, Butler PC. Beta-cell deficit and increased beta-cell apoptosis in humans with type 2 diabetes. Diabetes. 2003;52(1):102-110.

37. Weyer C, Bogardus C, Mott DM, Pratley RE. The natural history of insulin secretory dysfunction and insulin resistance in the pathogenesis of type 2 diabetes mellitus. J Clin Invest. 1999;104(6):787-794.

38. Weir GC. Non-insulin-dependent diabetes mellitus: interplay between B-cell inadequacy and insulin resistance. Am J Med. 1982;73(4):461-464.

39. Beagley J, Guariguata L, Weil C, Motala AA. Global estimates of undiagnosed diabetes in adults. Diabetes Res Clin Pract. 2014;103(2):150-160.

40. Alkayyali S, Lyssenko V. Genetics of diabetes complications. Mamm Genome. 2014;25(9-10):384-400.

41. Calles-Escandon J, Cipolla M. Diabetes and endothelial dysfunction: a clinical perspective. Endocr Rev. 2001;22(1):36-52.

42. Kashihara N, Haruna Y, Kondeti VK, Kanwar YS. Oxidative stress in diabetic nephropathy. Curr Med Chem. 2010;17(34):4256-4269.

43. Rodriguez-Poncelas A, Garre-Olmo J, Franch-Nadal J, Diez-Espino J, Mundet-Tuduri X, Barrot-De la Puente J, Coll-de Tuero G. Prevalence of chronic kidney disease in patients with type 2 diabetes in Spain: PERCEDIME2 study. BMC Nephrol. 2013;14(46):1-8.

44. Mason RM, Wahab NA. Extracellular matrix metabolism in diabetic nephropathy. J Am Soc Nephrol. 2003;14(5):1358-1373.

45. Anil Kumar P, Welsh GI, Saleem MA, Menon RK. Molecular and cellular events mediating glomerular podocyte dysfunction and depletion in diabetes mellitus. Front Endocrinol (Lausanne). 2014;5:1-10.

46. Chade AR. Renal vascular structure and rarefaction. Compr Physiol. 2013;3(2):817-831.

47. Adler AI, Stevens RJ, Manley SE, Bilous RW, Cull CA, Holman RR. Development and progression of nephropathy in type 2 diabetes: the United Kingdom Prospective Diabetes Study (UKPDS 64). Kidney Int. 2003;63(1):225-232.

48. Maezawa Y, Takemoto M, Yokote K. Cell biology of diabetic nephropathy: Roles of endothelial cells, tubulointerstitial cells and podocytes. J Diabetes Investig. 2015;6(1):3-15.

49. Suh JH, Miner JH. The glomerular basement membrane as a barrier to albumin. Nat Rev Nephrol. 2013;9(8):470-477.

50. Brennan E, McEvoy C, Sadlier D, Godson C, Martin F. The genetics of diabetic nephropathy. Genes (Basel). 2013;4(4):596-619.

51. Mogensen CE. Microalbuminuria predicts clinical proteinuria and early mortality in maturity-onset diabetes. N Engl J Med. 1984;310(6):356-360.

52. Valmadrid CT, Klein R, Moss SE, Klein BE. The risk of cardiovascular disease mortality associated with microalbuminuria and gross proteinuria in persons with older-onset diabetes mellitus. Arch Intern Med. 2000;160(8):1093-1100.

53. Rossing P. Diabetic nephropathy: worldwide epidemic and effects of current treatment on natural history. Curr Diab Rep. 2006;6(6):479-483.

54. Gall MA, Hougaard P, Borch-Johnsen K, Parving HH. Risk factors for development of incipient and overt diabetic nephropathy in patients with non-insulin dependent diabetes mellitus: prospective, observational study. BMJ. 1997;314(7083):783-788.

55. Ravid M, Brosh D, Ravid-Safran D, Levy Z, Rachmani R. Main risk factors for nephropathy in type 2 diabetes mellitus are plasma cholesterol levels, mean blood pressure, and hyperglycemia. Arch Intern Med. 1998;158(9):998-1004.

56. Toeller M, Buyken A, Heitkamp G, Bramswig S, Mann J, Milne R, Gries FA, Keen H. Protein intake and urinary albumin excretion rates in the EURODIAB IDDM Complications Study. Diabetologia. 1997;40(10):1219-1226.

57. Association AD. Diabetic Nephropathy (Position Statement). Diabetes Care. 2002;25(Suppl 1):S85-S89.

58. Zelmanovitz T, Gerchman F, Balthazar AP, Thomazelli FC, Matos JD, Canani LH. Diabetic nephropathy. Diabetol Metab Syndr. 2009;1(1):1-17.

59. Mauer SM, Steffes MW, Ellis EN, Sutherland DE, Brown DM, Goetz FC. Structural-functional relationships in diabetic nephropathy. J Clin Invest. 1984;74(4):1143-1155.

60. Kolset SO, Reinholt FP, Jenssen T. Diabetic nephropathy and extracellular matrix. J Histochem Cytochem. 2012;60(12):976-986.

61. Kimmelstiel P, Wilson C. Intercapillary Lesions in the Glomeruli of the Kidney. Am J Pathol. 1936;12(1):83-98.

62. Ayodele OE, Alebiosu CO, Salako BL. Diabetic nephropathy--a review of the natural history, burden, risk factors and treatment. J Natl Med Assoc. 2004;96(11):1445-1454.

63. Hostetter TH, Troy JL, Brenner BM. Glomerular hemodynamics in experimental diabetes mellitus. Kidney Int. 1981;19(3):410-415.

64. Hostetter TH, Rennke HG, Brenner BM. The case for intrarenal hypertension in the initiation and progression of diabetic and other glomerulopathies. Am J Med. 1982;72(3):375-380.

65. Muskiet MH, Smits MM, Morsink LM, Diamant M. The gut-renal axis: do incretin-based agents confer renoprotection in diabetes? Nat Rev Nephrol. 2014;10(2):88-103.

66. Mogensen CE, Christensen CK, Vittinghus E. The stages in diabetic renal disease. With emphasis on the stage of incipient diabetic nephropathy. Diabetes. 1983;32 Suppl 2:64-78.

67. Schlondorff D. The glomerular mesangial cell: an expanding role for a specialized pericyte. FASEB J. 1987;1(4):272-281.

68. Schlondorff D, Banas B. The mesangial cell revisited: no cell is an island. J Am Soc Nephrol. 2009;20(6):1179-1187.

69. Ghayur MN, Krepinsky JC, Janssen LJ. Contractility of the Renal Glomerulus and Mesangial Cells: Lingering Doubts and Strategies for the Future. Med Hypotheses Res. 2008;4(1):1-9.

70. Tsilibary EC. Microvascular basement membranes in diabetes mellitus. J Pathol. 2003;200(4):537-546.

71. Jefferson JA, Shankland SJ, Pichler RH. Proteinuria in diabetic kidney disease: a mechanistic viewpoint. Kidney Int. 2008;74(1):22-36.

72. Furness PN. Extracellular matrix and the kidney. J Clin Pathol. 1996;49(5):355-359.

73. Bonnans C, Chou J, Werb Z. Remodelling the extracellular matrix in development and disease. Nat Rev Mol Cell Biol. 2014;15(12):786-801.

74. Visse R, Nagase H. Matrix metalloproteinases and tissue inhibitors of metalloproteinases: structure, function, and biochemistry. Circ Res. 2003;92(8):827-839.

75. Xu X, Xiao L, Xiao P, Yang S, Chen G, Liu F, Kanwar YS, Sun L. A glimpse of matrix metalloproteinases in diabetic nephropathy. Curr Med Chem. 2014;21(28):3244-3260.

76. Lawrence DA. Latent-TGF-beta: an overview. Mol Cell Biochem. 2001;219(1-2):163-170.

77. Lee HS. Mechanisms and consequences of TGF-ss overexpression by podocytes in progressive podocyte disease. Cell Tissue Res. 2012;347(1):129-140.

78. Koli K, Saharinen J, Hyytiainen M, Penttinen C, Keski-Oja J. Latency, activation, and binding proteins of TGF-beta. Microsc Res Tech. 2001;52(4):354-362.

79. Leehey DJ, Singh AK, Alavi N, Singh R. Role of angiotensin II in diabetic nephropathy. Kidney Int Suppl. 2000;77:S93-98.

80. Weigert C, Sauer U, Brodbeck K, Pfeiffer A, Haring HU, Schleicher ED. AP-1 proteins mediate hyperglycemia-induced activation of the human TGF-beta1 promoter in mesangial cells. J Am Soc Nephrol. 2000;11(11):2007-2016.

81. Weigert C, Brodbeck K, Klopfer K, Haring HU, Schleicher ED. Angiotensin II induces human TGF-beta 1 promoter activation: similarity to hyperglycaemia. Diabetologia. 2002;45(6):890-898.

82. Hill CS. The Smads. Int J Biochem Cell Biol. 1999;31(11):1249-1254.

83. Miyazono K, ten Dijke P, Heldin CH. TGF-beta signaling by Smad proteins. Adv Immunol. 2000;75:115-157.

84. Miyazono K. Positive and negative regulation of TGF-beta signaling. J Cell Sci. 2000;113 (Pt 7):1101-1109.

85. Korchynskyi O, ten Dijke P. Identification and functional characterization of distinct critically important bone morphogenetic protein-specific response elements in the Id1 promoter. J Biol Chem. 2002;277(7):4883-4891.

86. Wolf G, Chen S, Ziyadeh FN. From the periphery of the glomerular capillary wall toward the center of disease: podocyte injury comes of age in diabetic nephropathy. Diabetes. 2005;54(6):1626-1634.

87. Li JJ, Kwak SJ, Jung DS, Kim JJ, Yoo TH, Ryu DR, Han SH, Choi HY, Lee JE, Moon SJ, Kim DK, Han DS, Kang SW. Podocyte biology in diabetic nephropathy. Kidney Int Suppl. 2007;72(106):S36-S42.

88. Satchell SC. The glomerular endothelium emerges as a key player in diabetic nephropathy. Kidney Int. 2012;82(9):949-951.

89. Tryggvason K, Patrakka J, Wartiovaara J. Hereditary proteinuria syndromes and mechanisms of proteinuria. N Engl J Med. 2006;354(13):1387-1401.

90. Satchell SC, Braet F. Glomerular endothelial cell fenestrations: an integral component of the glomerular filtration barrier. Am J Physiol Renal Physiol. 2009;296(5):F947-F956.

91. Roselli S, Heidet L, Sich M, Henger A, Kretzler M, Gubler MC, Antignac C. Early glomerular filtration defect and severe renal disease in podocin-deficient mice. Mol Cell Biol. 2004;24(2):550-560.

92. George B, Holzman LB. Signaling from the podocyte intercellular junction to the actin cytoskeleton. Semin Nephrol. 2012;32(4):307-318.

93. Mundel P, Shankland SJ. Podocyte biology and response to injury. J Am Soc Nephrol. 2002;13(12):3005-3015.

94. Qiu W, Zhou Y, Jiang L, Fang L, Chen L, Su W, Tan R, Zhang CY, Han X, Yang J. Genipin inhibits mitochondrial uncoupling protein 2 expression and ameliorates podocyte injury in diabetic mice. PLoS One. 2012;7(7):1-9.

95. Kriz W, Shirato I, Nagata M, LeHir M, Lemley KV. The podocyte's response to stress: the enigma of foot process effacement. Am J Physiol Renal Physiol. 2013;304(4):F333-F347.

96. Rippin JD, Barnett AH, Bain SC. Cost-effective strategies in the prevention of diabetic nephropathy. Pharmacoeconomics. 2004;22(1):9-28.

97. Lewis EJ, Hunsicker LG, Bain RP, Rohde RD. The effect of angiotensin-converting-enzyme inhibition on diabetic nephropathy. The Collaborative Study Group. N Engl J Med. 1993;329(20):1456-1462.

98. Ravid M, Brosh D, Levi Z, Bar-Dayan Y, Ravid D, Rachmani R. Use of enalapril to attenuate decline in renal function in normotensive, normoalbuminuric patients with type 2 diabetes mellitus - A randomized, controlled trial. Ann Intern Med. 1998;128(12):982-988.

99. Huang Y, Zhou Q, Haaijer-Ruskamp FM, Postma MJ. Economic evaluations of angiotensin-converting enzyme inhibitors and angiotensin II receptor blockers in type 2 diabetic nephropathy: a systematic review. BMC Nephrol. 2014;15(15):1-17.

100. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747-753.

101. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet. 2014;95(1):5-23.

102. Wang Q, Lu Q, Zhao H. A review of study designs and statistical methods for genomic epidemiology studies using next generation sequencing. Front Genet. 2015;6:1-12.

103. Ku CS, Cooper DN, Polychronakos C, Naidoo N, Wu M, Soong R. Exome sequencing: dual role as a discovery and diagnostic tool. Ann Neurol. 2012;71(1):5-14.

104. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet. 2010;11(6):415-425.

105. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends Genet. 2014;30(9):418-426.

106. Raje N, Soden S, Swanson D, Ciaccio CE, Kingsmore SF, Dinwiddie DL. Utility of next generation sequencing in clinical primary immunodeficiencies. Curr Allergy Asthma Rep. 2014;14(10):468-489.

107. Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ, Hart E, Suner MM, Landrum MJ, Aken B, Ayling S, Baertsch R, Fernandez-Banet J, Cherry JL, Curwen V, Dicuccio M, Kellis M, Lee J, Lin MF, Schuster M, Shkeda A, Amid C, Brown G, Dukhanina O, Frankish A, Hart J, Maidak BL, Mudge J, Murphy MR, Murphy T, Rajan J, Rajput B, Riddick LD, Snow C, Steward C, Webb D, Weber JA, Wilming L, Wu W, Birney E, Haussler D, Hubbard T, Ostell J, Durbin R, Lipman D. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome Res. 2009;19(7):1316-1323.

108. Wu L, Schaid DJ, Sicotte H, Wieben ED, Li H, Petersen GM. Case-only exome sequencing and complex disease susceptibility gene discovery: study design considerations. J Med Genet. 2015;52(1):10-16.

109. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakkaloglu A, Ozen S, Sanjad S, Nelson-Williams C, Farhi A, Mane S, Lifton RP. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci U S A. 2009;106(45):19096-19101.

110. Prada CE, Gonzaga-Jauregui C, Tannenbaum R, Penney S, Lupski JR, Hopkin RJ, Sutton VR. Clinical utility of whole-exome sequencing in rare diseases: Galactosialidosis. Eur J Med Genet. 2014;57(7):339-344.

111. Anderson MW, Schrijver I. Next generation DNA sequencing and the future of genomic medicine. Genes (Basel). 2010;1(1):38-69.

112. Raffan E, Semple RK. Next generation sequencing--implications for clinical practice. Br Med Bull. 2011;99:53-71.

113. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74(12):5463-5467.

114. Maxam AM, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci U S A. 1977;74(2):560-564.

115. Finishing the euchromatic sequence of the human genome. Nature. 2004;431(7011):931-945.

116. Reis-Filho JS. Next-generation sequencing. Breast Cancer Res. 2009;11 Suppl 3:S12-S19.

117. Monticone S, Else T, Mulatero P, Williams TA, Rainey WE. Understanding primary aldosteronism: Impact of next generation sequencing and expression profiling. Mol Cell Endocrinol. 2014;399:311-320.

118. Mardis ER. The impact of next-generation sequencing technology on genetics. Trends Genet. 2008;24(3):133-141.

119. Frese KS, Katus HA, Meder B. Next-generation sequencing: from understanding biology to personalized medicine. Biology (Basel). 2013;2(1):378-398.

120. Hui P. Next generation sequencing: chemistry, technology and applications. Top Curr Chem. 2014;336:1-18.

121. Zhu P, He L, Li Y, Huang W, Xi F, Lin L, Zhi Q, Zhang W, Tang YT, Geng C, Lu Z, Xu X. OTG-snpcaller: an optimized pipeline based on TMAP and GATK for SNP calling from ion torrent data. PLoS One. 2014;9(5):1-9.

122. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC, Gibbs RA. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One. 2012;7(11):1-12.

123. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics. 2012;13:1-13.

124. Merriman B, Rothberg JM. Progress in ion torrent semiconductor chip based sequencing. Electrophoresis. 2012;33(23):3397-3417.

125. Mardis ER. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet. 2008;9:387-402.

126. Dolled-Filhart MP, Lee M, Jr., Ou-Yang CW, Haraksingh RR, Lin JC. Computational and bioinformatics frameworks for next-generation whole exome and genome sequencing. ScientificWorldJournal. 2013;2013:1-10.

127. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-2079.

128. Bromberg Y. Building a genome analysis pipeline to predict disease risk and prevent disease. J Mol Biol. 2013;425(21):3993-4005.

129. Day-Williams AG, Zeggini E. The effect of next-generation sequencing technology on complex trait research. Eur J Clin Invest. 2011;41(5):561-567.

130. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156-2158.

131. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. Nucleic Acids Res. 2002;30(17):3894-3900.

132. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248-249.

133. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4(7):1073-1081.

134. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13):3812-3814.

135. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46(3):310-315.

136. QIAGEN. DNeasy® Blood & Tissue Handbook. 2006:9-11.

137. Technologies L. Ion AmpliSeq™ DNA and RNA Library Preparation. 2014:15-34.

138. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics. 2010;26(16):2069-2070.

139. Chen Y, Cunningham F, Rios D, McLaren WM, Smith J, Pritchard B, Spudich GM, Brent S, Kulesha E, Marin-Garcia P, Smedley D, Birney E, Flicek P. Ensembl variation resources. BMC Genomics. 2010;11:1-16.

140. Paila U, Chapman BA, Kirchner R, Quinlan AR. GEMINI: integrative exploration of genetic variation and genome annotations. PLoS Comput Biol. 2013;9(7):1-8.

141. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308-311.
142. Consortium TEP. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science. 2004;306(5696):636-640.
143. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42(Database issue):D980-D985.
144. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65.
145. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, Kang HM, Jordan D, Leal SM, Gabriel S, Rieder MJ, Abecasis G, Altshuler D, Nickerson DA, Boerwinkle E, Sunyaev S, Bustamante CD, Bamshad MJ, Akey JM. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science. 2012;337(6090):64-69.
146. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27-30.
147. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 2005;15(7):901-913.
148. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A. Human Protein Reference Database--2009 update. Nucleic Acids Res. 2009;37(Database issue):D767-D772.
149. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: integrating information about genes, proteins and diseases. Trends Genet. 1997;13(4):163.
150. Smith RN, Aleksic J, Butano D, Carr A, Contrino S, Hu F, Lyne M, Lyne R, Kalderimis A, Rutherford K, Stepan R, Sullivan J, Wakeling M, Watkins X, Micklem G.

InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. Bioinformatics. 2012;28(23):3163-3165.

151. Lane L, Argoud-Puy G, Britan A, Cusin I, Duek PD, Evalet O, Gateau A, Gaudet P, Gleizes A, Masselot A, Zwahlen C, Bairoch A. neXtProt: a knowledge platform for human proteins. Nucleic Acids Res. 2012;40(Database issue):D76-D83.

152. Cooper DN, Stenson PD, Chuzhanova NA. The Human Gene Mutation Database (HGMD) and its exploitation in the study of mutational mechanisms. Curr Protoc Bioinformatics. 2006;Chapter 1:Unit 1 13.

153. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res. 2004;32(Database issue):D115-D119.

154. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010;42(4):348-354.

155. Korbie DJ, Mattick JS. Touchdown PCR for increased specificity and sensitivity in PCR amplification. Nat Protoc. 2008;3(9):1452-1456.

156. Lewis CM. Genetic association studies: design, analysis and interpretation. Brief Bioinform. 2002;3(2):146-153.

157. Mahajan R, Mishra B. Using Glycated Hemoglobin HbA1c for diagnosis of Diabetes mellitus: An Indian perspective. Int J Biol Med Res. 2011;2(2):508-512.

158. Doria A. Genetics of diabetes complications. Curr Diab Rep. 2010;10(6):467-475.

159. Technologies A. Bioanalyzer Applications for Next-Gen Sequencing: Updates and Tips. 2011:4-49.

160. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014;15(2):121-132.

161. Dogan RI, Getoor L, Wilbur WJ, Mount SM. SplicePort--an interactive splice-site analysis tool. Nucleic Acids Res. 2007;35(Web Server issue):W285-W291.

162. Desmet FO, Hamroun D, Lalande M, Collod-Beroud G, Claustres M, Beroud C. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res. 2009;37(9):1-14.

163. Reese MG, Eeckman FH, Kulp D, Haussler D. Improved splice site detection in Genie. J Comput Biol. 1997;4(3):311-323.

164. Freund M, Asang C, Kammler S, Konermann C, Krummheuer J, Hipp M, Meyer I, Gierling W, Theiss S, Preuss T, Schindler D, Kjems J, Schaal H. A novel approach to describe a U1 snRNA binding site. Nucleic Acids Res. 2003;31(23):6963-6975.

165. Brendel V, Xing L, Zhu W. Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus. Bioinformatics. 2004;20(7):1157-1169.

166. Huang C, Kim Y, Caramori ML, Fish AJ, Rich SS, Miller ME, Russell GB, Mauer M. Cellular basis of diabetic nephropathy: II. The transforming growth factor-beta system and diabetic nephropathy lesions in type 1 diabetes. Diabetes. 2002;51(12):3577-3581.

167. Heydemann A, Ceco E, Lim JE, Hadhazy M, Ryder P, Moran JL, Beier DR, Palmer AA, McNally EM. Latent TGF-beta-binding protein 4 modifies muscular dystrophy in mice. J Clin Invest. 2009;119(12):3703-3712.

168. van den Bergen JC, Hiller M, Bohringer S, Vijfhuizen L, Ginjaar HB, Chaouch A, Bushby K, Straub V, Scoto M, Cirak S, Humbertclaude V, Claustres M, Scotton C, Passarelli C, Lochmuller H, Muntoni F, Tuffery-Giraud S, Ferlini A, Aartsma-Rus AM, Verschuuren JJ, t Hoen PA, Spitali P. Validation of genetic modifiers for Duchenne muscular dystrophy: a multicentre study assessing SPP1 and LTBP4 variants. J Neurol Neurosurg Psychiatry. 2015;86(10):1060-1065.

169. Ha H, Hwang IA, Park JH, Lee HB. Role of reactive oxygen species in the pathogenesis of diabetic nephropathy. Diabetes Res Clin Pract. 2008;82 Suppl 1:S42-S45.

170. Dando I, Fiorini C, Pozza ED, Padroni C, Costanzo C, Palmieri M, Donadelli M. UCP2 inhibition triggers ROS-dependent nuclear translocation of GAPDH and autophagic cell death in pancreatic adenocarcinoma cells. Biochim Biophys Acta. 2013;1833(3):672-679.

171. Giacco F, Brownlee M. Oxidative stress and diabetic complications. Circ Res. 2010;107(9):1058-1070.

172. Brand MD, Esteves TC. Physiological functions of the mitochondrial uncoupling proteins UCP2 and UCP3. Cell Metab. 2005;2(2):85-93.

173. Souza BM, Michels M, Sortica DA, Boucas AP, Rheinheimer J, Buffon MP, Bauer AC, Canani LH, Crispim D. Polymorphisms of the UCP2 Gene Are Associated with

Glomerular Filtration Rate in Type 2 Diabetic Patients and with Decreased UCP2 Gene Expression in Human Kidney. PLoS One. 2015;10(7):1-15.

174. Lu MK, Gong XG, Guan KL. mTOR in podocyte function: is rapamycin good for diabetic nephropathy? Cell Cycle. 2011;10(20):3415-3416.

175. Yasuda M, Tanaka Y, Kume S, Morita Y, Chin-Kanasaki M, Araki H, Isshiki K, Araki S, Koya D, Haneda M, Kashiwagi A, Maegawa H, Uzu T. Fatty acids are novel nutrient factors to regulate mTORC1 lysosomal localization and apoptosis in podocytes. Biochim Biophys Acta. 2014;1842(7):1097-1108.

176. Godel M, Hartleben B, Herbach N, Liu S, Zschiedrich S, Lu S, Debreczeni-Mor A, Lindenmeyer MT, Rastaldi MP, Hartleben G, Wiech T, Fornoni A, Nelson RG, Kretzler M, Wanke R, Pavenstadt H, Kerjaschki D, Cohen CD, Hall MN, Ruegg MA, Inoki K, Walz G, Huber TB. Role of mTOR in podocyte function and diabetic nephropathy in humans and mice. J Clin Invest. 2011;121(6):2197-2209.

177. Foster KG, Acosta-Jaquez HA, Romeo Y, Ekim B, Soliman GA, Carriere A, Roux PP, Ballif BA, Fingar DC. Regulation of mTOR complex 1 (mTORC1) by raptor Ser863 and multisite phosphorylation. J Biol Chem. 2010;285(1):80-94.

178. Inoki K, Mori H, Wang J, Suzuki T, Hong S, Yoshida S, Blattner SM, Ikenoue T, Ruegg MA, Hall MN, Kwiatkowski DJ, Rastaldi MP, Huber TB, Kretzler M, Holzman LB, Wiggins RC, Guan KL. mTORC1 activation in podocytes is a critical step in the development of diabetic nephropathy in mice. J Clin Invest. 2011;121(6):2181-2196.

179. Sengupta S, Peterson TR, Sabatini DM. Regulation of the mTOR complex 1 pathway by nutrients, growth factors, and stress. Mol Cell. 2010;40(2):310-322.

180. Oshiro N, Yoshino K, Hidayat S, Tokunaga C, Hara K, Eguchi S, Avruch J, Yonezawa K. Dissociation of raptor from mTOR is a mechanism of rapamycin-induced inhibition of mTOR function. Genes Cells. 2004;9(4):359-366.

181. Abu Seman N, He B, Ojala JR, Wan Mohamud WN, Ostenson CG, Brismar K, Gu HF. Genetic and biological effects of sodium-chloride cotransporter (SLC12A3) in diabetic nephropathy. Am J Nephrol. 2014;40(5):408-416.

182. Tanaka N, Babazono T, Saito S, Sekine A, Tsunoda T, Haneda M, Tanaka Y, Fujioka T, Kaku K, Kawamori R, Kikkawa R, Iwamoto Y, Nakamura Y, Maeda S. Association of solute carrier family 12 (sodium/chloride) member 3 with diabetic

nephropathy, identified by genome-wide analyses of single nucleotide polymorphisms. Diabetes. 2003;52(11):2848-2853.

183. Kim JH, Shin HD, Park BL, Moon MK, Cho YM, Hwang YH, Oh KW, Kim SY, Lee HK, Ahn C, Park KS. SLC12A3 (solute carrier family 12 member [sodium/chloride] 3) polymorphisms are associated with end-stage renal disease in diabetic nephropathy. Diabetes. 2006;55(3):843-848.

184. Ng DP, Nurbaya S, Choo S, Koh D, Chia KS, Krolewski AS. Genetic variation at the SLC12A3 locus is unlikely to explain risk for advanced diabetic nephropathy in Caucasians with type 2 diabetes. Nephrol Dial Transplant. 2008;23(7):2260-2264.

185. Mathieson PW. The podocyte cytoskeleton in health and in disease. Clin Kidney J. 2012;5(6):498-501.

186. Nakatani S, Kakehashi A, Ishimura E, Yamano S, Mori K, Wei M, Inaba M, Wanibuchi H. Targeted proteomics of isolated glomeruli from the kidneys of diabetic rats: sorbin and SH3 domain containing 2 is a novel protein associated with diabetic nephropathy. Exp Diabetes Res. 2011;2011:1-11.

187. Garg P, Verma R, Nihalani D, Johnstone DB, Holzman LB. Neph1 cooperates with nephrin to transduce a signal that induces actin polymerization. Mol Cell Biol. 2007;27(24):8698-8712.

188. Goley ED, Welch MD. The ARP2/3 complex: an actin nucleator comes of age. Nat Rev Mol Cell Biol. 2006;7(10):713-726.

189. Jaziri R, Aubert R, Roussel R, Emery N, Maimaitiming S, Bellili N, Miot A, Saulnier PJ, Travert F, Hadjadj S, Marre M, Fumeron F. Association of ADIPOQ genetic variants and plasma adiponectin isoforms with the risk of incident renal events in type 2 diabetes. Nephrol Dial Transplant. 2010;25(7):2231-2237.

190. Kankova K, Stejskalova A, Pacal L, Tschoplova S, Hertlova M, Krusova D, Izakovicova-Holla L, Beranek M, Vasku A, Barral S, Ott J. Genetic risk factors for diabetic nephropathy on chromosomes 6p and 7q identified by the set-association approach. Diabetologia. 2007;50(5):990-999.

191. Cooke Bailey JN, Palmer ND, Ng MC, Bonomo JA, Hicks PJ, Hester JM, Langefeld CD, Freedman BI, Bowden DW. Analysis of coding variants identified from exome sequencing resources for association with diabetic and non-diabetic nephropathy in African Americans. Hum Genet. 2014;133(6):769-779.

192. Freedman BI, Hicks PJ, Sale MM, Pierson ED, Langefeld CD, Rich SS, Xu J, McDonough C, Janssen B, Yard BA, van der Woude FJ, Bowden DW. A leucine repeat in the carnosinase gene CNDP1 is associated with diabetic end-stage renal disease in European Americans. Nephrol Dial Transplant. 2007;22(4):1131-1135.

193. Ahluwalia TS, Lindholm E, Groop LC. Common variants in CNDP1 and CNDP2, and risk of nephropathy in type 2 diabetes. Diabetologia. 2011;54(9):2295-2302.

194. Tong Z, Yang Z, Patel S, Chen H, Gibbs D, Yang X, Hau VS, Kaminoh Y, Harmon J, Pearson E, Buehler J, Chen Y, Yu B, Tinkham NH, Zabriskie NA, Zeng J, Luo L, Sun JK, Prakash M, Hamam RN, Tonna S, Constantine R, Ronquillo CC, Sadda S, Avery RL, Brand JM, London N, Anduze AL, King GL, Bernstein PS, Watkins S, Jorde LB, Li DY, Aiello LP, Pollak MR, Zhang K. Promoter polymorphism of the erythropoietin gene in severe diabetic eye and kidney complications. Proc Natl Acad Sci U S A. 2008;105(19):6998-7003.

195. Moczulski DK, Grzeszczak W, Gawlik B. Role of hemochromatosis C282Y and H63D mutations in HFE gene in development of type 2 diabetes and diabetic nephropathy. Diabetes Care. 2001;24(7):1187-1191.

196. Alkayyali S, Lajer M, Deshmukh H, Ahlqvist E, Colhoun H, Isomaa B, Rossing P, Groop L, Lyssenko V. Common variant in the HMGA2 gene increases susceptibility to nephropathy in patients with type 2 diabetes. Diabetologia. 2013;56(2):323-329.

197. Buraczynska M, Swatowski A, Buraczynska K, Dragan M, Ksiazek A. Heat-shock protein gene polymorphisms and the risk of nephropathy in patients with Type 2 diabetes. Clin Sci (Lond). 2009;116(1):81-86.

198. Blakemore AI, Cox A, Gonzalez AM, Maskil JK, Hughes ME, Wilson RM, Ward JD, Duff GW. Interleukin-1 receptor antagonist allele (IL1RN*2) associated with nephropathy in diabetes mellitus. Hum Genet. 1996;97(3):369-374.

199. Cooke JN, Bostrom MA, Hicks PJ, Ng MC, Hellwege JN, Comeau ME, Divers J, Langefeld CD, Freedman BI, Bowden DW. Polymorphisms in MYH9 are associated with diabetic nephropathy in European Americans. Nephrol Dial Transplant. 2012;27(4):1505-1511.

200. Kuricova K, Tanhauserova V, Pacal L, Bartakova V, Brozova L, Jarkovsky J, Kankova K. NOS3 894G>T polymorphism is associated with progression of kidney

disease and cardiovascular morbidity in type 2 diabetic patients: NOS3 as a modifier gene for diabetic nephropathy? Kidney Blood Press Res. 2013;38(1):92-98.

201. McKnight AJ, Currie D, Patterson CC, Maxwell AP, Fogarty DG. Targeted genome-wide investigation identifies novel SNPs associated with diabetic nephropathy. Hugo J. 2009;3(1-4):77-82.

202. Neves AL, Mohammedi K, Emery N, Roussel R, Fumeron F, Marre M, Velho G. Allelic variations in superoxide dismutase-1 (SOD1) gene and renal and cardiovascular morbidity and mortality in type 2 diabetic subjects. Mol Genet Metab. 2012;106(3):359-365.

203. Buraczynska M, Baranowicz-Gaszczyk I, Borowicz E, Ksiazek A. TGF-beta1 and TSC-22 gene polymorphisms and susceptibility to microvascular complications in type 2 diabetes. Nephron Physiol. 2007;106(4):69-75.

204. Ahluwalia TS, Lindholm E, Groop L, Melander O. Uromodulin gene variant is associated with type 2 diabetic nephropathy. J Hypertens. 2011;29(9):1731-1734.

205. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011;89(1):82-93.

206. Vetter SW, Indurthi VS. Moderate glycation of serum albumin affects folding, stability, and ligand binding. Clin Chim Acta. 2011;412(23-24):2105-2116.

207. Tamura Y, Adachi H, Osuga J, Ohashi K, Yahagi N, Sekiya M, Okazaki H, Tomita S, Iizuka Y, Shimano H, Nagai R, Kimura S, Tsujimoto M, Ishibashi S. FEEL-1 and FEEL-2 are endocytic receptors for advanced glycation end products. J Biol Chem. 2003;278(15):12613-12617.

208. Peppa M, Vlassara H. Advanced glycation end products and diabetic complications: a general overview. Hormones (Athens). 2005;4(1):28-37.

209. Sourris KC, Harcourt BE, Penfold SA, Yap FY, Morley AL, Morgan PE, Davies MJ, Baker ST, Jerums G, Forbes JM. Modulation of the cellular expression of circulating advanced glycation end-product receptors in type 2 diabetic nephropathy. Exp Diabetes Res. 2010;2010:1-9.

210. Forbes JM, Cooper ME, Oldfield MD, Thomas MC. Role of advanced glycation end products in diabetic nephropathy. J Am Soc Nephrol. 2003;14(8 Suppl 3):S254-S258.

211. Yamagishi S, Matsui T. Advanced glycation end products, oxidative stress and diabetic nephropathy. Oxid Med Cell Longev. 2010;3(2):101-108.
212. English WR, Velasco G, Stracke JO, Knauper V, Murphy G. Catalytic activities of membrane-type 6 matrix metalloproteinase (MMP25). FEBS Lett. 2001;491(1-2):137-142.
213. Maeda S, Haneda M, Guo B, Koya D, Hayashi K, Sugimoto T, Isshiki K, Yasuda H, Kashiwagi A, Kikkawa R. Dinucleotide repeat polymorphism of matrix metalloproteinase-9 gene is associated with diabetic nephropathy. Kidney Int. 2001;60(4):1428-1434.
214. Fragiadaki M, Ikeda T, Witherden A, Mason RM, Abraham D, Bou-Gharios G. High doses of TGF-beta potently suppress type I collagen via the transcription factor CUX1. Mol Biol Cell. 2011;22(11):1836-1844.
215. Clout NJ, Tisi D, Hohenester E. Novel fold revealed by the structure of a FAS1 domain pair from the insect cell adhesion molecule fasciclin I. Structure. 2003;11(2):197-203.
216. Appella E, Weber IT, Blasi F. Structure and function of epidermal growth factor-like regions in proteins. FEBS Lett. 1988;231(1):1-4.
217. Sohail A, Sun Q, Zhao H, Bernardo MM, Cho JA, Fridman R. MT4-(MMP17) and MT6-MMP (MMP25), A unique set of membrane-anchored matrix metalloproteinases: properties and expression in cancer. Cancer Metastasis Rev. 2008;27(2):289-302.
218. Harada R, Berube G, Tamplin OJ, Denis-Larose C, Nepveu A. DNA-binding specificity of the cut repeats from the human cut-like protein. Mol Cell Biol. 1995;15(1):129-140.
219. Ramdzan ZM, Nepveu A. CUX1, a haploinsufficient tumour suppressor gene overexpressed in advanced cancers. Nat Rev Cancer. 2014;14(10):673-682.
220. Pufall MA, Graves BJ. Autoinhibitory domains: modular effectors of cellular regulation. Annu Rev Cell Dev Biol. 2002;18:421-462.
221. Burkhard P, Stetefeld J, Strelkov SV. Coiled coils: a highly versatile protein folding motif. Trends Cell Biol. 2001;11(2):82-88.
222. Ohta M, Matsui K, Hiratsu K, Shinshi H, Ohme-Takagi M. Repression domains of class II ERF transcriptional repressors share an essential motif for active repression. Plant Cell. 2001;13(8):1959-1968.

CHAPTER 6| Appendices

Appendix A

Table A1. Type 2 diabetes susceptibility genes (adapted from (9)).

Gene	Chromosome	Odds ratio	RAF	Type of study	Function and probable mechanism
ADAMTS9	3	1.05-1.09	0.68-0.81	MA	Metalloproteinase/ Insulin action
ADCY5	3	1.12	0.78	MA	Adenylyl cyclases/ Insulin action
ANK1	8	1.09	0.76	MA	Cell stability/ β-cell function
ANKRD55	5	1.08	0.7	MA	Insulin action
ANKS1A	6	1.11	0.91	GWAS	Pathway regulator/ Unknown
ARAP1	11	1.08-1.14	0.81-0.88	GWAS, MA	Actin cytoskeleton modulator/ β-cell function
BCAR1	16	1.12	0.89	MA	Docking protein/ β-cell function
BCL11A	2	1.08-1.09	0.46	MA	Zinc finger/ β-cell function
BCL2	18	1.09	0.64	GWAS	Cell death regulator/Unknown
CAMK1D	10	1.07-1.11	0.18	LA, MA	Protein kinase/ β-cell function
CAPN10	2	1.09-1.18	0.73-0.96	MA	Calpain cysteine protease/ Insulin action
CDKAL1	6	1.10-1.20	0.27-0.31	GWAS, MA	β-cell function
CDKN2A	9	1.19-1.20	0.82-0.83	GWAS	Cyclin-dependent kinase inhibitor/ β-cell function
CENTD2	11	1.08-1.13	0.81-0.88	GWAS	β-cell function

Gene	Chromosome	Odds ratio	RAF	Type of study	Function and probable mechanism
CHCHD9	9	1.11-1.20	0.93	MA	Unknown
CILP2	19	1.13	0.08	MA	Unknown
DGKB	7	1.04-1.06	0.47-0.54	MA	Diacylglycerol kinase/ Insulin action
DUSP9	X	1.09-1.27	0.12-0.77	MA	Phosphatase
FOLH1	11	1.10	0.09	GWAS	Transmembrane glycoprotein/ Unknown
FTO	16	1.06-1.27	0.38-0.41	GWAS, MA	Metabolic regulator/ Insulin action
GATAD2A	19	1.12	0.08	GWAS	Transcriptional repressor/ Unknown
GCK	7	1.07	0.20	MA	Glucokinase/ Insulin action
GCKR	2	1.06-1.09	0.59-0.62	MA	Glucokinase regulator/ Insulin action
GIPR	19	1.10	0.27	GWAS	G-protein coupled receptor/ Unknown
GRB14	2	1.07	0.60	GCS, MA	Adapter protein/ Insulin action
HFE	6	1.12	0.29	MA	Membrane protein/ Unknown
HHEX	10	1.12-1.13	0.53-0.60	LA, MA	Transcriptional repressor/ Intracellular insulin degradation/ Motor protein
HMG20A	15	1.08	0.68	GCS, MA	Chromatin-associated protein/ Unknown
HMGA1	6	1.34-15.8	0.10	GCS	Transcriptional regulator/ Insulin action

Gene	Chromosome	Odds ratio	RAF	Type of study	Function and probable mechanism
HMGA2	12	1.10-1.20	0.09-0.10	MA	Transcriptional regulator
HNF1A	12	1.07-1.14	0.77-0.85	MA	Pancreatic and liver transcriptional activator
HNF1B	17	1.08-1.17	0.47-0.51	GCS, MA	Transcription factor/ β-cell function
IGF2BP2	3	1.14	0.29-0.32	GWAS, MA	Binding protein/ β-cell function
IRS1	2	1.09-1.12	0.64-0.67	GCS, MA	Insulin signaling element/ Insulin action
JAZF1	7	1.10	0.52	MA	Zing finger/ β-cell function
KCNJ11	11	1.09-1.14	0.37-0.47	GCS, MA	Potassium channel/ β-cell function
KCNQ1	11	1.08-1.23	0.44	GWAS	Potassium channel/ β-cell function
KLF14	7	1.07-1.10	0.55	MA	Transcription factor/Insulin action
KLHDC5	12	1.10	0.80	MA	Mitotic progression and cytokinesis/Unknown
LAMA1	18	1.13	0.38	GWAS	Cellular migration mediator/Unknown
MC4R	18	1.08	0.27	MA	G-protein-coupled receptor/Unknown
MTNR1B	11	1.05-1.08	0.28-0.30	GWAS, MA	Melatonin receptor/ β-cell function
NOTCH2	1	1.06-1.13	0.10-0.11	MA	Membrane receptor
PPARG	3	1.11-1.17	0.85-0.88	GCS, MA	Nuclear receptor/ Insulin action
PRC1	15	1.07-1.10	0.22	MA	Cytokinesis regulador
PROX1	1	1.07	0.50	MA	Homeobox transcription factor/ Insulin action
PTPRD	9	1.57	0.10	GWAS	Protein tyrosine phosphatase

Gene	Chromosome	Odds ratio	RAF	Type of study	Function and probable mechanism
RBMS1	2	1.08-1.11	0.79-0.83	MA	DNA modulator/ Insulin action
SLC2A2	3	1.06	0.74	GWAS	Glucose sensor/β-cell function
SLC30A8	8	1.11-1.18	0.65-0.70	GWAS, MA	Zinc efflux transporter/β-cell function
SREBF1	17	1.07	0.38	GWAS	Lipid transcriptional regulator/ Unknown
SRR	17	1.28	0.69	GWAS	Serine racemase
TCF7L2	10	1.31-1.71	0.26-0.30	GWAS, LA, MA	Participates in the Wnt signaling pathway/β-cell function
THADA	2	1.15	0.90	MA	Thyroid adenoma-associated protein/β-cell function
TH/INS	11	1.14	0.39	GWAS	Catecholamine synthesis/ Unknown
TLE1	9	1.07	0.57	MA	Transcriptional corepressor/ Unknown
TP53INP1	8	1.06-1.11	0.48	MA	Proapoptotic protein/Unknown
TSPAN8	12	1.06-1.09	0.27-0.71	MA	Cell surface glycoprotein/ β-cell function
WFS1	4	1.10-1.13	0.60-0.73	GCS	Transmembrane protein/ β-cell function
ZBED3	5	1.08-1.16	0.26	MA	Zinc finger/ β-cell function
ZFAND6	15	1.01-1.11	0.60-0.72	MA	Zinc finger/ β-cell function
ZMIZ1	10	1.08	0.52	MA	Transcriptional regulator/ Unknown

DNA: deoxyribonucleic acid; GCS: gene candidate studies; GWAS: genome-wide association study; LA: linkage analysis; MA: meta-analysis; RAF: **r**isk allele frequency.

Appendix B

Table B1. Characterization of the control group.

Exomes in study	Age (years)	Disease duration (years)	HbA1c (%)	Sex
21	58	9	14.3	Female
22	71	22	9.1	Female
23	63	16	6.5	Female
29	62	13	7.6	Male
34	65	24	8.9	Female
37	64	22	7.7	Male
38	74	5	10.5	Female
44	57	33	11.0	Female
46	62	13	7.7	Female
47	65	23	10.9	Female
50	58	15	9.0	Female
134	67	2	10.6	Male
136	70	22	8.5	Male
142	49	2	13.2	Male
155	64	2	11.2	Male
158	49	11	8.4	Male
161	45	3	12.6	Male
166	75	29	6.8	Female
168	63	13	8.3	Male

HbA1c: glycated hemoglobin.

Table B2. Characterization of the case group.

Exomes in study	Age (years)	Disease duration (years)	HbA1c (%)	Sex
2	70	23	9.3	Male
3	70	23	8.1	Female
4	72	14	11.0	Male
7	70	33	7.1	Female
16	74	5	5.2	Male
17	68	17	8.6	Male
27	77	19	9.8	Female
39	65	23	8.5	Male
45	70	20	9.3	Male
114	57	15	5.7	Male
116	67	2	12.6	Male
123	67	12	7.0	Male
132	56	2	5.4	Female
140	75	8	13.5	Female
149	66	32	12.8	Male
151	62	17	8.6	Male
164	67	16	6.2	Female

HbA1c: glycated hemoglobin.

Appendix C

Table C1. Coverage analysis for each exome.

Exomes in study	Coverage Analysis metrics			
	Mapped Reads (number)	**Reads on Target (%)**	**Mean Depth (x)**	**Uniformity (%)**
2	40,259,514	93.96	111.30	67.52
3	29,810,924	93.00	75.33	68.34
4	18,516,510	91.97	46.17	68.69
7	40,455,509	92.12	107.90	61.78
16	47,570,535	94.50	138.70	91.87
17	39,713,195	94.35	115.20	92.34
21	42,355,561	94.97	124.30	93.21
22	45,374,642	94.77	132.80	92.92
23	42,919,052	94.57	123.70	92.99
27	37,295,987	93.52	105.60	90.32
29	42,658,040	95.00	124.70	91.93
34	60,004,978	93.78	168.40	93.10
37	49,030,450	96.22	152.30	93.68
38	41,249,229	95.20	120.70	92.38
39	39,685,963	95.25	116.60	92.36
44	41,630,753	95.21	121.30	92.87
45	41,059,571	95.28	119.40	92.79
46	40,323,607	95.57	118.90	91.95
47	43,292,335	95.33	127.20	93.30
50	38,643,370	95.24	110.30	91.92
114	31,222,625	95.44	83.14	87.90
116	27,080,919	95.74	72.44	90.33
123	37,021,319	95.49	99.20	85.67
132	27,478,730	96.31	80.32	87.55
134	48,586,140	96.67	152.90	93.65
136	47,905,254	96.35	149.90	91.88
140	47,762,220	96.43	150.30	93.00
142	47,476,003	96.13	147.20	92.19
149	46,238,152	95.90	144.30	92.79
151	47,599,369	96.08	147.10	89.40
155	49,609,805	96.67	148.70	92.07
158	47,531,704	96.09	140.70	92.26
161	42,348,690	96.54	132.60	92.41
164	45,105,472	96.31	139.90	89.62
166	32,978,117	96.85	103.70	58.15
168	41,626,664	96.52	130.00	56.66

Table C2. Total of genetic variants by type and number of homozygous and heterozygous for each variant type by exome.

Exomes in study	SNP			INS			DEL			MNP		
	HM	HT	Total	HM	HT	Total	HM	HT	Total	HM	HT	Total
2	14,481	34,946	49,427	405	2,136	2,541	483	2,962	3,445	15	426	441
3	12,912	34,312	47,224	379	4,224	4,603	476	4,402	4,878	16	422	438
4	12,754	29,188	41,942	303	1,937	2,240	906	3,431	4,337	13	287	300
7	14,275	31,840	46,115	471	2,878	3,349	725	2,780	3,505	2	19	21
16	18,787	29,765	48,552	490	798	1,288	636	1,355	1,991	19	42	61
17	18,411	30,718	49,129	572	834	1,406	537	1,291	1,828	14	31	45
21	18,782	29,907	48,689	524	801	1,325	582	1,260	1,842	16	22	38
22	18,365	30,974	49,339	495	863	1,358	672	1,361	2,033	16	41	57
23	18,131	31,260	49,391	517	855	1,372	582	1,494	2,076	11	38	49
27	18,193	30,098	48,291	498	822	1,320	580	1,321	1,901	16	32	48
29	18,382	30,644	49,026	547	864	1,411	569	1,387	1,956	13	54	67
34	18,149	31,580	49,729	592	976	1,568	568	1,695	2,263	27	98	125
37	18,644	31,319	49,963	609	1,033	1,642	520	1,124	1,644	155	354	509
38	18,732	30,525	49,257	493	794	1,287	683	1,390	2,073	11	34	45
39	17,946	31,413	49,359	489	770	1,259	615	1,401	2,016	14	28	42
44	18,349	30,928	49,277	488	842	1,330	547	1,330	1,877	23	32	55
45	18,550	30,380	48,930	556	916	1,472	549	1,435	1,984	21	59	80
46	18,182	31,045	49,227	549	858	1,407	536	1,199	1,735	10	47	57
47	18,478	30,946	49,424	518	799	1,317	629	1,324	1,953	18	41	59
50	17,979	30,667	48,646	517	907	1,424	516	1,396	1,912	16	53	69
114	17,052	28,486	45,538	630	1,325	1,955	786	1,763	2,549	87	242	329
116	16,984	29,006	45,990	639	1,259	1,898	841	1,730	2,571	91	258	349
123	17,442	28,309	45,751	610	1,284	1,894	861	1,672	2,533	114	228	342
132	19,237	25,579	44,816	541	872	1,413	515	878	1,393	151	396	547
134	18,792	31,196	49,988	580	1,002	1,582	478	1,158	1,636	152	368	520
136	18,851	30,825	49,676	552	969	1,521	535	1,151	1,686	145	359	504
140	18,712	31,407	50,119	605	1,028	1,633	510	1,124	1,634	144	342	486
142	18,839	30,744	49,583	593	1,046	1,639	491	1,077	1,568	149	361	510
149	18,664	31,446	50,110	580	1,033	1,613	512	1,104	1,616	156	351	507
151	19,996	27,767	47,763	578	926	1,504	547	994	1,541	152	366	518
155	18,794	31,175	49,969	625	1,005	1,630	488	1,239	1,727	150	410	560
158	18,469	31,582	50,051	618	1,070	1,688	518	1,135	1,653	157	387	544
161	18,735	30,747	49,482	568	945	1,513	533	1,159	1,692	148	393	541
164	18,162	31,706	49,868	582	960	1,542	494	1,150	1,644	144	392	536
166	13,801	21,241	35,042	387	645	1,032	351	631	982	124	298	422
168	13,524	21,971	35,495	386	704	1,090	337	668	1,005	125	300	425

DEL: deletion; HM: homozygouss; HT: heterozygous; INS: insertion; MNP: multiple nucleotide polymorphism; SNP: single nucleotide polymorphism.

Appendix D

Table D1. List of the filtered common genetic variants obtained in the statistical analysis of the 36 exomes.

Gene	rs ID	Ref. allele	Alt. allele	Type of variant
ABCC11	rs61739606	T	A	Missense
ACTRT2	rs3795263	G	A	Missense
ADAM8	rs3008326	G	A	Missense
AIM1L	rs11247919	C	T	Splice
AMDHD1	rs7955450	A	G	Missense
ARHGEF26	rs13096373	T	C	Missense
ARPC2	rs10169718	A	G	Splice
ASB16-AS1	rs7220138	C	G	Splice
ASMTL	None	T	C	Splice
ASTN2	rs72765708	C	T	Missense
BACE1	rs490460	C	A	Splice
BFAR	rs11546303	T	G	Missense
C3	rs1047286	G	A	Missense
C17orf53	rs227584	A	C	Missense
C17orf102	rs58529418	C	G	Missense
CCNB1IP1	rs1132644	G	T	Splice
CD207	rs10489990	G	A	Missense
CDH4	rs6142884	A	G	Missense
CEP44	rs4695918	G	A	Missense
CEP89	rs10411735	C	T	Splice
CKMT2	rs545	G	A	3'-UTR
CLEC7A	None	A	C	Stop gained
CNN2	rs2304260	G	A	Missense
COL17A1	rs17116350	T	C	Missense
CYBRD1	rs10455	G	A	Missense
DENND2A	rs2293177, rs386564002	C	T	Missense
DHX35	rs3752302	C	T	Missense
DNAJB13	rs72982975	G	A	Missense
EIF2AK3	rs867529	G	C	Missense
EPB41	rs12070152	A	G	Intron
ERCC2	rs13181	T	G	Missense
	rs1799793	C	T	Missense
FAHD1	rs3743853	G	A	Missense
FAM86A	rs12928528	G	C	Missense
FAM179A	rs11127202	A	G	Missense
FANCI	rs17803620	C	T	Missense
	rs2283432	G	C	Missense
FBXO39	rs4796555	C	T	Missense

Gene	rs ID	Ref. allele	Alt. allele	Type of variant
FIGLA	rs7566476	C	G	Missense
GABRR1	rs1186902	T	C	Missense
GALNT14	rs2288101, rs386563728	G	T	Missense
GGT5	rs762276	G	A	Missense
GPR98	rs2366777	G	T	Missense
	rs4916684	G	A	Missense
	rs2247870	G	A	Missense
GREB1	rs2304402	G	A	Missense
HEATR4	rs8014577	T	C	Splice
HNRNPC	rs8016099	A	C	Intron
IFNL1	rs30461, rs386578910	A	G	Missense
IL15	rs2857261	A	G	Splice
INMT	rs4720015	T	G	Missense
IRF6	rs7552506	G	C	Splice
ITGA10	rs2274616	G	A	Missense
ITIH4	rs13072536	A	T	Missense
JSRP1	rs80043033	C	G	Missense
KRT3	rs3887954	G	C	Missense
KRTAP10-6	rs233303	G	T	Stop gained
LAMTOR4	rs3736591	A	G	Splice
LDHAL6B	rs3825937, rs386587320	T	C	Missense
LGALS9	rs361497	G	A	Missense
LMO7	rs7986131	T	C	Missense
LTBP4	rs1051303	A	G	Missense
	rs1131620	A	G	Missense
MERTK	rs7604639	G	A	Missense
	rs2230515	A	G	Missense
MESDC2	rs56314660	G	A	Intron
MFSD6	rs386624005, rs9646748	A	G	Missense
MLN	rs2281820	A	G	Missense
MYO16	rs157024	A	G	Missense
NACA2	rs17531723	C	T	Missense
	rs61739273	G	C	Missense
NAV2	rs6483617	G	A	Missense
NCKAP5	rs12611515	C	T	Missense
	rs12691830	A	G	Missense
OBSCN	rs453140	C	T	Missense
	rs1188697	G	A	Missense
	rs11810627	C	T	Missense
	rs435776	G	A	Missense
	rs437129	G	A	Missense
	rs1188710	C	G	Missense
OR2T3	rs139993642	C	T	Missense
OR4N4	rs62006710	A	G	Missense
	rs62006708	G	C	Missense

Gene	rs ID	Ref. allele	Alt. allele	Type of variant
OR7C1	rs73004304	C	G	Missense
OR10R2	rs3820678	G	A	Missense
OR10V1	rs499033	T	C	Missense
OR51F1	rs17324812	T	C	Missense
P2RX3	rs2276038	C	T	Missense
PBXIP1	rs2061690	C	T	Missense
PCGF2	rs1138349	G	A	Stop gained
PKHD1L1	rs1673408	C	A	Missense
PRSS56	rs2853447	A	G	Splice
RAET1E	rs9383583	C	T	Missense
RAPGEF3	rs2074533	T	C	Splice
RP1L1	rs55642448	C	T	Missense
RP5-966M1.6	rs2071041	T	G	Splice
RPF1	rs2292191	A	G	Missense
RPTOR	rs2589156	G	A	Splice
SEPT9	rs34587622	C	T	Missense
SERPINB8	rs3826616	A	G	Missense
SIGLEC10	rs9304711	G	A	Missense
SLC12A3	rs2304483	T	C	Splice
SLC12A7	rs11951420	G	A	Splice
SLC22A1	rs628031	A	G	Missense
SLCO2A1	rs34550074	C	T	Missense
SPHKAP	rs16824284	T	C	Intron
SRRM2	rs16824284	C	A	Missense
STAC	rs112766131	G	A	5’-UTR
	rs113135397	C	T	5’-UTR
STARD3	rs16824284	G	A	Missense
SULF2	rs56218501	C	T	Missense
SYT6	rs72690087	G	T	Intron
TAF4B	rs74947492	G	C	Missense
TAS1R2	rs34447754	G	C	Missense
	rs35874116	T	C	Missense
TCOF1	rs15251	C	T	Missense
	rs1136103	C	G	Missense
TEC	rs2271173	G	A	Splice
TMC6	rs2613522	A	G	Splice
TMPRSS4	rs2276122	A	G	Splice
TUBB6	rs11267036	G	C	Splice
UBASH3B	rs12790613	G	A	Missense
UBE3B	rs7298565	G	A	Missense
UBR4	rs3762396	A	G	Splice
UCP2	rs660339	G	A	Missense
UNC79	rs4905081	G	A	Missense
VANGL1	rs4839469	G	A	Missense

Gene	rs ID	Ref. allele	Alt. allele	Type of variant
VN1R4	rs112711591, rs74429916	G	A	Missense
WDYHV1	rs4416808	T	A	Splice
ZFR2	rs2240231	G	A	Splice
	rs2240235	G	A	Missense
ZMYM5	rs41292167	T	C	Missense
	rs9579718	T	C	Missense
ZNF419	rs1135692, rs8108040	G	C	Splice
ZNF860	rs13087612	T	C	Missense

Alt.: altered; Ref.: reference; UTR: untranslated region.

Table D2. Impact prediction for the splice region variants in the common variants approach.

Prediction Tool	Reference Sequence > Sequence with the variant		
	rs2589156 (*RPTOR*)	**rs2304483 (*SLC12A3*)**	**rs10169718 (*ARPC2*)**
SplicePort: An Interactive Splice Site Analysis Tool (161)*	1.97 > loss of the splice region	0.04 > - 0.17	0.46 > - 0.14
HSF (162)**	91.61 > 62.66	87.98 > 88.59	65.74 > 65.91
Analyzer Splice Tool***	95.04 > 83.12	79.19 > 77.24	The variant is outside of the region analyzed by the prediction tool
NNSplice (163)****	0.99 > 0.89	The prediction tool does not recognize the region as a splice region	0.99 > 0.99
HBond Score Web-Interface (164)*****	19.90 > 14.70	The variant is outside of the region analyzed by the prediction tool	17.20 > 15.30
USD SplicePredictor Online Service (165)******	0.99 > 0.94	The prediction tool does not recognize the region as a splice region	0.90 > loss of the splice region

*spliceport.cbcb.umd.edu
**www.umd.be/HSF3/index.html
***ibis.tau.ac.il/ssat/SpliceSiteFrame.htm
****www.fruitfly.org/seq_tools/splice.html
*****www.uni-duesseldorf.de/rna/html/hbond_score.php
******bioservices.usd.edu/splicepredictor/

Table D3. List of diabetic nephropathy candidate genes and respective genetic variants for European type 2 diabetic individuals.

Gene	Genetic variant ID	Reference
ADIPOQ	rs17300539	(189)
	rs2241766	
AGER	2184A>G	(190)
APOL2	rs7285167	(191)
CNDP1	5-5 leucine repeat, also termed CNDP1 Mannheim	(192)
	rs2346061	(193)
CNDP2	rs7577	(193)
EPO	rs1617640	(194)
HFE	rs1799945	(195)
HMGA2	rs1531343	(196)
HSPA1A	rs1008438	(197)
	rs1043618	
IL1RN	86 bp variable number tandem repeat polymorphism in intron 2	(198)
LIMK2	rs61098917	(191)
	rs3747154	
MYH9	rs4821480	(199)
	rs4281481	
	rs2032487	
	rs3752462	
NOS3	894G>T	(200)
OR2AK2	rs4478844	(191)
PLXND1	rs2285372	(201)
	rs2301572	
RAET1L	rs1543547	(201)
SOD1	rs1041740	(202)
TGFB1	869T>C	(203)
UMOD	rs13333226	(204)

bp: base-pairs.

Appendix E

Table E1. List of the statistically significant genes accumulating rare variants in the 36 exomes.

Gene	Number of rare variants accumulated	Number of individuals with the accumulated rare variants
ABCC6	3	2
ADAMTS8	3	6
ADD1	2	2
AMH	2	2
ANKAR	2	2
APOA4	3	3
APOC4	2	2
APPBP2	1	2
ARFGEF2	2	2
ARHGAP44	2	2
ASCC2	3	3
BPG116M5.17	2	4
BPIFC	2	3
BRCA2	2	2
C1orf129	3	3
C6orf1321	1	2
C16orf7	1	2
C17orf64	1	2
CAST	2	2
CCDC64B	2	2
CD93	2	2
CECR2	5	4
CELSR1	3	4
CENPE	2	2
CERCAM	4	3
CFB	2	4
CFH	3	4
CLCN2	2	2
CNTNAP5	1	2
COL6A2	2	2
CUX1	2	2
CYP2W1	2	2
DDX60	1	2
DENND1C	3	3
DNAH2	17	13
DNAI1	5	5
DOK3	2	2
DOPEY1	3	4

Gene	Number of rare variants accumulated	Number of individuals with the accumulated rare variants
EYS	3	3
FAM65A	3	3
FAM129C	2	2
FAM160A1	5	4
FAM193A	4	3
FBN3	11	12
FBRSL1	4	5
FCRL1	3	4
FUT11	2	2
GANAB	2	2
GBGT1	4	4
GRN	1	2
GTF2F1	1	2
HABP2	4	6
HIRA	1	2
HOXC6	1	3
HUWE1	2	2
IDI2	1	3
IGSF21	2	2
ITGA9	2	2
ITGB7	4	4
IQCE	2	2
KCNH5	2	2
KCNQ5	2	2
KIAA0195	2	2
KIAA1244	1	2
LRRC37A3	4	4
LTK	3	4
METRNL	2	2
MMP25	2	2
NEU4	3	4
NFYA	2	2
NRAP	4	4
NT5DC2	5	5
OBP2A	2	2
OR11A1	3	4
OSGIN1	3	3
OTUD7A	2	2
PAM	2	2
PARP10	2	2
PC	2	2
PDIA2	6	7
PEG3	5	7

Gene	Number of rare variants accumulated	Number of individuals with the accumulated rare variants
PGBD2	2	1
PI15	2	2
PIK3C2A	3	3
PIK3C2B	3	3
PKD1	14	12
PLA1A	2	2
PLOD2	2	2
PMS1	2	3
POLD1	4	5
PTPN13	4	4
RAPGEF2	2	2
RELN	4	5
REV3L	2	2
RPL36	1	2
SDHD	2	2
SLC22A13	3	3
SLITRK5	2	2
SPG7	1	2
SRD5A1	2	2
ST7	1	2
STAB1	9	10
TANC1	3	3
TARM1	2	2
TBCC	2	2
TCHH	2	2
TEX14	2	2
TMC3	3	3
TMEM43	1	2
TMEM173	3	3
TMPRSS7	5	3
TOP3A	1	3
TP53I13	2	2
TXNDC16	3	3
VPS8	2	2
VWA5B1	4	4
WDR25	2	3
WDR36	3	3
YLPM1	4	4
ZNF134	2	2
ZNF197	2	2
ZNF208	7	4
ZNF280B	3	4

Gene	Number of rare variants accumulated	Number of individuals with the accumulated rare variants
ZNF343	2	3
ZNF407	4	3
ZNF416	2	2
ZNF680	1	2
ZNF844	3	3

Table E2. Impact prediction for the splice region variants accumulated in the genes obtained from the rare variants approach.

Prediction Tool	Reference Sequence > Sequence with the variant		
	STAB1 **(rs371042844)**	***STAB1*** **(rs199636230)**	***STAB1*** **(rs143836348)**
SplicePort: An Interactive Splice Site Analysis Tool (161)*	2.38 > 2.54	The prediction tool does not recognize the region as a splice region	0.31 > 0.11
HSF (162)**	The prediction tool does not recognize the region as a splice region	85.60 > 77.92	67.90 > 66.05
Analyzer Splice Tool***	The variant is outside of the region analyzed by the prediction tool	81.48 > 75.89	The variant is outside of the region analyzed by the prediction tool
NNSplice (163)****	1.00 > 0.99	0.57 > 0.69	0.94 > 0.91
HBond Score Web-Interface (164)*****	17.10 > 17.10	The variant is outside of the region analyzed by the prediction tool	The variant is outside of the region analyzed by the prediction tool
USD SplicePredictor Online Service (165)******	1.00 > 1.00	0.86 > 0.92	The prediction tool does not recognize the region as a splice region

*spliceport.cbcb.umd.edu
**www.umd.be/HSF3/index.html
***ibis.tau.ac.il/ssat/SpliceSiteFrame.htm
****www.fruitfly.org/seq_tools/splice.html
*****www.uni-duesseldorf.de/rna/html/hbond_score.php
******bioservices.usd.edu/splicepredictor/